Eine Neufassung des reliable-change Index mit einer Anwendung in der Essstörungsforschung

Von der Gemeinsamen Naturwissenschaftlichen Fakultät
der Technischen Universität Carolo-Wilhelmina
zu Braunschweig
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr.rer.nat.)

genehmigte

Dissertation

von:
Peter Malewski
aus Osterode am Harz

1. Referent: Prof. Dr. W. Schulz
2. Referent: PD. Dr. B. Jäger
eingereicht am: 15.1.2004
mündliche Prüfung (Disputation) am: 23.6.2004

(2004)

Vorveröffentlichungen der Dissertation

Teilergebnisse aus dieser Arbeit wurden mit Genehmigung der Gemeinsamen Naturwissenschaft-lichen Fakultät, vertreten durch den Mentor der Arbeit, in folgenden Beiträgen vorab veröffentlicht:

Publikationen

P. Malewski, W. Schulz & B. Jäger (2003) Eine begriffliche Neufassung des reliable-change Begriffs mit einer auf forschungslogischen Konsequenzen reflektierenden Anwendung im Bereich der Essstörungsforschung. 54. Jahrestagung des Deutschen Kollegiums für Psychosomatische Medizin (DKPM). Abstrakt veröffentlich in: *Psychotherapie Psychosomatik Medizinische* Psychologie, 53, Seite: 11.

Tagungsbeiträge

P. Malewski, W. Schulz & B. Jäger A conceptual revised version of the reliable-change index with an application in eating-disorder research. 34. Tagung der Society for Psychotherapy Research (SPR), Weimar (2003).

meinen Eltern

Danksagung

An der Entstehung der vorliegenden Arbeit waren viele Personen beteiligt, denen ich an dieser Stelle herzlich danken möchte. Danken möchte ich insbesondere meinen beiden Betreuern Prof. Dr. Wolfgang Schulz und PD Dr. Burkard Jäger, die die Arbeit über Jahre hinweg im besten Sinne kritisch begleitet haben. Ganz besonders möchte ich auch Dr. Hans Kordy danken, der zu der Arbeit angeregt hat und gerade in der Anfangsphase grundlegende Ideen beisteuerte.

Weiterhin möchte ich allen ehemaligen Kollegen der Forschungsstelle für Psychotherapie Stuttgart für die sehr kollegiale Zusammenarbeit und die bereichernden Diskussionen danken. Insbesondere danke ich Dr. Jens Oehlschlägel, dass er mir an einigen Punkten mit seinem scharfen methodischen Verstand kritisch zur Seite stand.

Auch möchte ich mich sehr herzlich bei den Mitarbeitern der Abteilung Psychosomatik und Psychotherapie der Medizinischen Hochschule Hannover bedanken. Insbesondere bei Prof. Dr. Friedhelm Lamprecht und Prof. Dr. Gerhard Schmid-Ott, die mir durch kollegialen und wissenschaftlichen Rat zur Seite standen. Bei Frau Felden und Frau Welke möchte ich mich für die Mühe bei der Fehlerkorrektur bedanken. Weiterhin möchte ich mich recht herzlich bei Frau Behrens für die vielfältige Unterstützung, die sie mir gerade in der Endphase der Arbeit entgegenbrachte, bedanken. Auch möchte ich meinem Lehranalytiker, Dr. Michael Kögler, für seine emotionale Unterstützung und, diesmal nicht im methodischen Sinn gemeinten, analytischen Sachverstand danken.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Die vorliegende Arbeit kann in zwei Ebenen eingeteilt werden. Als erstes wird eine bestimmte Methodik der Psychotherapieforschung modifiziert und diese Modifizierung angewandt. Zweitens wird dieser Vorgang hinsichtlich seiner forschungslogischen Implikationen reflektiert. Die folgende Einleitung spiegelt diesen Sachverhalt wieder. Sie führt zum einen in die Überlegungen ein, das Konzept der Klinischen Signifikanz (CS) zu modifizieren zum anderen umreißt sie die abstraktere, forschungslogische Relexionsebene. Dabei verweisen die Abschnitte dieses Kapitels auf die zugehörigen, folgenden Kapitel, in denen die hier angeschnittene Problematik detaillierter erläutert wird.

1.1 Überblick zur Modifikation des reliable-change (RC) Indizes

Im Gegensatz zu anderen in der Psychotherapieforschung verbreiteten Erfolgsmaßen, ist ein Satz wie: ‚Die Studie X zeigte, dass 30% der Patienten, die vorher krank waren, jetzt wieder gesund sind und dass durch die Therapie eine bedeutsame Änderung erreicht wurde' auch einem Psychotherapiepatienten verständlich. Vielleicht würde, falls man diesen Patienten weiter fragte, was denn ‚krank' eigentlich bedeute, er mit ‚anormal' antworten; jedoch hätte er wohl für den Begriff der bedeutsamen Änderung keine annähernd richtige Antwort parat oder würde auf die Stärke der Änderung rekurieren. Dabei wäre diese Antwort nicht falsch, jedoch käme nur ein Forscher zu der zutreffenden Antwort, dass eben die Menge an Veränderungen, die über dem durch den Messfehler bedingten Bereich liegen, als bedeutsame Änderung bezeichnet wird.

Diese Sequenz mag den Grundgedanken des Konzepts der Clinical significance (CS) kurz beleuchtet haben (siehe Kapitel 2 auf Seite 5). Es handelt sich bei dem CS-Begriff um einen Schluss auf klinische Relevanz einer Änderung, d.h. es handelt sich immer um ein Vergleich zweier Merkmalsausprägungen zu zwei Zeitpunkten. Ziel ist jedoch nicht die kausale Attribution, dass die Änderung wg. einem anderen Einfluss zu Stande kam, sondern um die reine Feststellung der Existenz einer Änderung, d.h. es interessiert nicht das *weswegen* sondern, *dass* sich etwas geändert hat.

Der Schluss auf eine bedeutsame Änderung (CS) ist dann erfüllt, wenn zwei Urteile zutreffen: (1) Der Patient muss vorher krank, d.h. sich in einem Bereich der Merkmalsausprägung befinden, der nicht im Bereich der Normalität liegt, und später gesund sein (CSK) *und* (2) die Änderung der Merkmalsausprägung muss bedeutsam sein, d.h. über dem durch den Messfehler bedingten Bereich liegen (RC). Formal kann dies wie folgt

ausgedrückt werden:

$$CSK \wedge RC \Rightarrow CS \tag{1.1}$$

Hier ist also auf eine begriffliche Überschneidung aufmerksam zu machen: Da von den Originalautoren sowohl das erste Kriterium wie auch das ganze Konzept als Clinical Significance bezeichnet wird, ist es notwendig, diese beiden zu unterscheiden[1]. Hierbei wird CSK das CS-Kriterium bezeichnen und CS das Konzept als ganzes (eine ausführliche Diskussion möglicher Definitionen findet sich in Abschnitt 2.4 auf Seite 22). Für beide Teilbereiche des Schlusses müssen verschiedene Voraussetzungen erfüllt sein, die sich aus verschiedenen anderen, untergeorneten Begrifflichkeiten ableiten. Im ersten Fall ist dies der Normalitätsbegriff, im zweiten Fall der Begriff der Zuverlässigkeit (Reliabilität).

Warum ist in diesem zusammengesetzten Index einmal der Normalitätsbegriff, dann jedoch der Reliabilitätsbegriff Voraussetzung? Es liegt nahe, auch das zweite Urteil, das den Schluss auf CSK zulässt (RC) ebenfalls durch den Normalitätsbegriff zu begründen. Es ergäbe sich mithin ein Konzept wie das der „Normalität der Veränderung“ (siehe Kapitel 6 auf Seite 50).

Dieser Gedanke des Vergleichs mit der Normalität der Veränderung liegt implizit dem Kontrollgruppendesign zugrunde. Verglichen werden hier behandelte Patienten mit unbehandelten Kontrollgruppenpatienten. Ist die Erfolgsrate in der Behandlungsgruppe höher als in der Kontrollgruppe, spricht man von einer erfolgreichen Therapie. Hier wird also mit der ‚Normalität der Veränderung‘ verglichen — ein anderer Name dafür ist auch die ‚Spontanheilungsquote‘[2].

In der vorliegenden Arbeit wird folglich zu dem zweiten Kriterium des clinical significance Ansatzes eine Erweiterung vorgeschlagen, die sich in Richtung des Begriffs Normalität der Veränderung interpretieren lässt. Dieser Index soll demnach nicht nur eine Korrektur des Messfehlers, sondern auch die normalerweise zu erwartende Variabilität berücksichtigen. Da Veränderung immer in der Zeit stattfindet, erweitert sich mithin das Konzept der klinischen Signifikanz um eine zeitliche Komponente: Je mehr Zeit für eine Veränderung zur Verfügung steht, desto höher die Rate der zu erwarten-

[1] Diese Begriffsverwirrung ist wie folgt erklärlich. Die Originalautoren sahen in dem RC-Kriterium nur eine Hilfskonstruktion; in ihrem Gedankengang war das CSK eigentlich identisch mit dem CS. Das RC-Kriteium sollte alleine eine statistische Absicherung leisten. In der Literatur findet man verschiedene Übersetzungen des englischen ‚clinical significance‘-Begriffs in das Deutsche. Im englischen Original steht ‚clinical significance‘ sowohl für das Konzept als ganzes als auch für den ersten Teil dieses. Analog dazu Schmitz (1997): er übersetzt clinical significance mit ‚Klinische und Statistische Signifikanz‘. Diese Übersetzung ist jedoch etwas ‚schief‘, da auch gerade der erste Teil des Indizes ‚statistisch‘ ist, freilich auf andere Weise als der zweite Teil. So beruht, wie schon beschrieben, der zweite Teil auf Begriffen wie der ‚Wahrscheinlichkeit des Messfehlers‘. Dagegen ließe sich jedoch argumentieren, dass der ‚reliable change‘ Teil zwar ebenso statistisch, jedoch viel weniger ‚klinisch signifikant‘ wäre. Auch leuchtet die weitere Verwendung des Anglizismus ‚signifikant‘ nicht ein, bietet sich doch die in der deutschen Sprache handliche Übersetzung ‚bedeutsam‘ an. In der vorliegenden Arbeit wird an der englischen Begrifflichkeit festgehalten. Zu den Bedeutungen der Begrifflichkeiten siehe auch Abschnitt 2.4 auf Seite 22.

[2] Anzumerken ist hier, dass nicht die Normalität der Veränderung per se addressiert wird, sondern die einer besonderen Subgruppe, nämlich die der Erkrankten. Ob die Veränderungsrate in dieser Subgruppe identisch ist mit der Gesamtgruppe aller Menschen, ist zumindest sehr fraglich (vgl. Abschnitt 9.2 auf Seite 86).

den, spontanen Veränderung und, analog dazu, je kleiner der Bereich in dem eine durch eine Behandlungsmaßnahme induzierte Veränderung noch als bedeutsam gelten kann. Was hat dieses neue Konzept für Auswirkungen auf die klinische Forschung? Eine Erprobung an einer großen, multizentrischen Studie soll hier das Konzept verdeutlichen und eine erste Schätzung der Korrektur ermöglichen (siehe Kapitel 6 auf Seite 50).

1.2 Überblick zur forschungslogisch-reflexiven Ebene der Arbeit

Es handelt sich also bei der vorliegenden Arbeit um eine Modifikation einer bereits bestehenden Methodik. Ergänzt werden soll diese methodische Zielsetzung durch eine Reflexion auf diesen Prozess (der Einführung einer neuen Begrifflichkeit). D.h. während der Begründung einer neuen Methodik ist jeweils dieser Prozess in seinem allgemeinen Charakter darzustellen. Es handelt sich also um eine Art doppeltes Spiel. Ergebnis ist der Aufweis der Wirksamkeit impliziter Begrifflichkeiten und die strukturelle Kennzeichung des Vorgangs als ganzen. Implizite Begrifflichkeiten sind neben der Distinktion von Schätzverfahren und schliessender Verfahren (Kapitel 3 auf Seite 24) insbesondere der Normalitätsbegriff (Kapitel 4 auf Seite 34), welcher den CS-Ansatz begründet. Der Forschungsprozess als solcher wird vor allem auf dem Hintergrund des Wissenschaftsbgriffs Heideggers (Kapitel 5 auf Seite 42) beleuchtet. Dieser ist inbesondere wegen seines reflexiven Charakters, dem Voraussetzen bestimmter Strukturen und der dadurch sich ergebenen konstitutiven Funktion, sowie wegen der allgemeinen Verbindung dieses Prozesses mit dem unsere Wirklichkeit bestimmenden Geschehens von Interesse. Dieser doppelte Prozess ist wiederum im Sinne Heideggers eingebettet in ein geschichtlich spezifisches Wahrheitsgeschehen. Im folgenden soll kurz auf diese Punkte eingegangen werden.

Die Kritik an dem Ritual des Null-Hypothesen-Test (NHT) ist die in der Literatur verbreitet und kann in ihrem historischen Gewordensein als „Perpetuieren einer verpassten Chance einer anfänglichen Entscheidung“ benannt werden. Hauptkritikpunkte sind die Unergiebigkeit dieses Rituals im Prozess wissenschaftlichen Fortschritts und die theoretische Nichtexistenz der angewandten Form statistischen Schließens (vgl. Kapitel 3 auf Seite 24). So hat sich insbesondere Gigerenzer (Gigerenzer, 2000; Gigerenzer & Murray, 1987; Gigerenzer, 1989a, 1998) immer wieder gegen dieses Ritual gewandt und sich für flexiblere Methoden zur Datenanalyse ausgesprochen. Der hier im Fokus stehende Index der klinischen Signifikanz kann als ein *Beispiel* dieser eher auf Schätzung ausgelegten Strategie verstanden werden.

Warum ist sowohl dem Forscher als auch dem Laien der Begriff der „Normalität“ so selbstverständlich? Es mag überraschen, dass „Normalität“ ein relativ junger Begriff ist. Das relativ spät (Ende des 19. Jahrhunderts) entstandene Diskursdispositiv „Normalität“ (Link, 1999) ist ein Beispiel für die bedeutungs*tragende* Funktion spezieller, mathematischer Vorstellungsinhalte. Die historische Analyse zeigt hier eine Verdrängung des Norm-Begriffs zugunsten eines Normalitätsbegriffs, also eine vorgestellte Einordnung des Einzelnen in eine (imaginäre) Verteilungsform, prototypisch der Gauß-Glocken-

Kurve. Dabei ist weniger die Bedeutsamkeit des Normalitäts-Begriffs in der psychologischen Forschung bemerkenswert, als die Rolle im alltäglichen Diskurs, ja sogar bis in die psychologischen Mechanismen des ‚Herrn Jedermann': eines einzelnen, *normalen* Menschen. Dabei ist nicht nur der statische Aspekt („das ist nicht-normal") mit seinen Selbststeuerungsmechanismen bedeutsam, sondern auch der dynamische: diese Selbstnormalisierung orientiert sich an symbolischen Verlaufskurven, wie sich dies in Begriffen wie des Umsteuerns, Abbremsens oder das explorierende testing of one's limits zeigt (vgl. Kapitel 4 auf Seite 34).

Als Beispiel der Bedeutsamkeit des „Normalitätdispositiv" für die psychologische Methodenlehre kann die Entwicklung des Konzepts der klinischen Signifikanz gewertet werden. Ein Teil des CSK Konzeptes gründet sich hierbei auf den nicht hinterfragten Vorstellungen zur Normalität. Dabei ist sicher ein Teil des „Einleucht"–effekts auf diese untergründige, kulturelle Begrifflichkeit rückführbar. D.h. da die Konzeptualisierung des CSK-Begriffs auf die diskurstragende Begrifflichkeit der Normalität implizit zurückgreift ist sie auf besondere Weise verständlich; —obwohl der Normalitätsbegriff selbst komplex ist, da er etwa von der Verteilung homogener Eigenschaften auf einer imaginären Dimension beruht, also auf verschiedenen Abstraktionen gründet. Dieser Einleuchteffekt vollzieht sich jedoch gleichsam ‚hinter dem Rücken' der Forscher.

In der ursprünglichen Bedeutung bezeichnete „Mathematik" (Τὰ ματέματα) das im Betrachten von Gegenständen im voraus Bekannte. Erst später engte sich die Bedeutung auf die Zahlen, als das sich am stärksten Aufdrängende, Selbstverständliche ein. In der heutigen Form der Wissenschaft spielt Mathematik, insbesondere die Statistik, eine besondere Rolle. Diese ist jedoch mehr als bloßes Mittel zum dem Zweck, innerhalb des wissenschaftlichen Diskurses richtiges festzustellen. Sie ist konstitutiv: sowohl als Kristallisationspunkt der logischen Struktur (vgl. auch die Diskussion um das statistische Testen vs. bedeutsamer Indizes) wie auch bedeutungsgebend und -tragend (vgl. die Diskussion um den Normalitätsbegriff). Dabei ist diese Konstellation eine späte Erscheinung im abendländischen Denken. Sie selbst beruht schon auf einer bestimmten Wirklichkeitsauffassung, d. h. dessen, was die Dinge für uns *sind*. Dass dies so ist, offenbart sich erst durch den Blick auf andere Zeitalter und deren Wissenschaft. Heute finden wir das Mathematische in seiner allgemeineren Form als Garant der *Strenge* der Wissenschaft. Dass dies nicht immer so war und vielleicht nicht immer so sein wird, ergibt sich daraus (vgl. Kapitel 5 auf Seite 42). Denn aus der Feststellung, dass die Wissenschaft eingebettet in ein Wahrheitsgeschehen ist, dass geschichtlich je ein anderes ist, ergibt sich, dass mit einer anderen Weise wie sich Wahrheit ereignet sich auch eine Wissenschaft ergibt.

2 Das clinical significance Konzept (CS)

Diese Arbeit fokussiert auf eine Modifikation des Konzepts der clinical significance. Ziel des CS-Konzeptes ist die Bestimmung des Erfolgs einer psychotherapeutischen Maßnahme (natürlich gelten diese Überlegungen auch für medizinische Eingriffe). Als erster Schritt wird, nach einigen Bemerkungen zu den möglichen Funktionen derartiger Maße, dieses Konzept dargestellt und analysiert. Die konkrete Modifikation findet sich im Kapitel 6 auf Seite 50.

Das CS-Konzept steht in einer Reihe anderer Erfolgsmaße, so etwa der verschiedenen Effektgrößen oder der odds-ratios. Diese Maße finden sich in verschiedenen Zusammenhängen wieder, so beispielsweise[1]:

- Innerhalb der Forschung:
  - *...um über die Effektivität eines Therapieverfahrens zu berichten*. Dies kann etwa im Rahmen von Studien, in denen ausschließlich die Behandlungsgruppe untersucht wird, um überhaupt die Durchführbarkeit eines Verfahrens zu demonstrieren oder in Studien, bei denen eine unbehandelte Kontrollgruppe mit untersucht wird, um die Wirksamkeit eines Verfahrens zu verdeutlichen, der Fall sein.
  - ...um *verschiedene* Therapieverfahren zu *vergleichen*. Hierbei kann es sich etwa um Studien handeln, die bestimmen wollen, welches Verfahren für welche Indikation geeigneter ist oder aber im Rahmen der Zulassung neuer Verfahren darstellen wollen, dass ein neues Verfahren mindestens genauso gut wie ein bereits zugelassenes ist (aber vielleicht kostengünstiger).
  - *...um Einflussgrößen auf den Erfolg eines Therapieverfahren zu bestimmen.*
  - ...oder um die Effektivität vieler Therapiearten *zusammenzufassen*.
- *Außerhalb* der Forschung:
  - ...in verschiedenen *Rechtfertigungszusammenhängen*: etwa gegenüber den Krankenkassen, gegenüber der Laienpresse etc. So kann etwa vor dem Hintergrund eines vermehrten Kostendruckes im Gesundheitswesen Angaben zur Effektivität von Behandlungsverfahren, insbesondere auch im Rahmen von Kosten-Nutzenrechnungen bedeutsam werden.

[1] Diese Liste ist sicher weder erschöpfend, noch nach einheitlichen Gesichtspunkten, wie etwa nach der Studienart, angeordnet, sondern dient ausschließlich zur Illustration

– ...und um außerhalb der Forschung Informationen zur Effektivität berichten zu können, etwa für angehende Therapiepatienten, die an der Frage, was sie sich denn eigentlich erhoffen können, interessiert sind..

So heterogen diese Zusammenhänge sind, in denen Informationen zur Effektivität von Psychotherapien relevant werden, so unterschiedlich ist auch die Nützlichkeit eines Indizes in diesen Zusammenhängen: So mag der Index X bei dem Vergleich zweier Therapieverfahren durchaus sinnvoll sein, gegenüber Y vielleicht sogar überlegen, in einem außerhalb der Forschung liegenden Rechtfertigungszusammenhang mag sich dies Verhältnis jedoch umkehren. Auch sind diese Bereiche nicht gänzlich unabhängig voneinander, gibt es doch auch Arbeiten, die man als Übersetzungen von Ergebnissen von einem in einen anderen Bereich verstehen kann: So etwa das Expertengutachten zur Psychotherapie, das den Forschungs- in ein Rechtfertigungszusammenhang verwandelt (Meyer et al., 1991). Ein Ziel muss im Rahmen der vorliegenden Arbeit jedoch, unter Ausklammerung derartiger Erschwernisse, die Einschätzung der Funktionalität des CS Indizes im Kontrast zu anderen, konkurrierenden Indizes innerhalb verschiedener Zusammenhänge sein. Wie schon in der Einleitung (vgl. Absch. 1 auf Seite 1) angerissen, ist die Nützlichkeit des clinical significance Ansatzes besonders im alltäglichen Diskurs evident. Die Nützlichkeit in anderen Zusammenhängen bleibt vorerst ein Diskussionspunkt, der nach der Vorstellung des Konzepts zu behandeln ist. Eine vorläufige Bewertung findet sich am Ende dieses Kapitels, eine abschließende Bewertung (auch der modifizierten Form dieses Ansatzes) findet sich in der Diskussion (vgl. Abschnitt 9.2 auf Seite 86).

Motivation zur Bildung des CS-Indizes war eine Analyse eben dieses Zusammenhanges. Jacobson, Folette, & Revensdorf (1984) rekurieren hier explizit auf den Alltagsdiskurs des klinisch Tätigen. Sie fragen explizit, was gängige Ergebnisse der Psychotherapieforschung im Diskurs des Klinikers bedeuten. D.h. die Motivation zur Bildung des CS-Begriffs war die Frage, welche Funktion eine aus der Mathematik stammende Begrifflichkeit im Diskurs derer hat, die an der mathematischen Konzeptualisierung ein Interesse zeigen.

2.1 Begriffliche Analyse des CS-Konzepts

Erfolg kann als Zielerreichung definiert werden. Aus klinischer Sicht ist das Ziel einer psychotherapeutischen Intervention eine positive Änderung des Gesundheitszustands, oder die **Heilung** (die altgriechische Übersetzung ist hier θεραπεία, also ‚Therapie')[1]. Von diesem Prinzip lassen sich zwei Momente abstrahieren:

1. die beiden Zustände: ‚gesund' und ‚krank'

2. und der Wechsel zwischen beiden.

[1]Dass bei einigen psychischen Erkrankungen dieses Ziel oft unrealistisch ist, d. h. eine Diskussion über das für ein Patientenkollektiv Erreichbare stattfindet, ändert nichts am Prinzip der Bewegung vom Zustand der Krankheit zu dem der Gesundheit.

Diese beiden Momente bilden die zwei Teile des CS-Ansatzes. Diese beiden Teile sind auf einer begrifflichen Ebene definiert, es fehlt jedoch eine in der Forschung verwendbare Form, d.h. es schließen sich verschiedene Konkretisierungen und die Übersetzung in eine mathematische Begrifflichkeit an. Durch diese Arbeit gewinnt das Konzept auch auf begrifflicher Ebene weitere Spezifizierungen, so dass das endgültige Konzept ein Ergebnis dieser Übersetzungsarbeit ist (in der zusammenfassenden Abbildung 6.1 auf Seite 50 ist dieser doppelseitige Vorgang wiedergegeben). Ausgehend von einem vagen Verständnis wird sich am Ende eine Definition des Begriffs der ‚clinical significance' ergeben.

2.2 Der erste Teil des Indizes: das clinical significance-Kriterium (CSK)

Im Folgenden wird versucht, ausgehend von einer inhaltlichen Analyse zu den konkreten methodischen Konzepten des clinical significance-Kriterium (CSK) hinzuleiten. Dementsprechend ‚mathematisiert' sich die Darstellung zunehmend.

2.2.1 Begriffliche Analyse und Ableitung des Indizes

Beide Teile des Indizes gehen auf unterschiedliche Weise von dem Begriff der Heilung aus. Der vorliegende erste Teil geht vor allem auf den statischen Aspekt dieses Begriffs ein. D.h. es handelt sich um die Zustände Gesundheit und Krankheit. Im folgenden ist also kurz auf den Gesundheitsbegriff einzugehen.

Jeder Arzt weiß: es gibt keinen absolut gesunden Menschen, jeder Mensch besitzt eine schwache Stelle, eine verdeckte Infektion etc. So hat etwa Antonovsky (1985, 1987) immer wieder betont: Gesundheit und Krankheit sind Kontinuen, d. h. es gibt nur ein mehr oder weniger, kein entweder-oder. Veranschaulicht werden kann dies durch die Vorstellung des Körpers als einer Zusammenstellung vieler Teile: eins kann bis zu einem gewissen Grade krank sein, wobei alle anderen funktionieren, d. h. durch die Zusammenstellung vieler Teile entsteht ein Kontinuum. Mithin könnte man jemandem, der von sich behauptete, er wäre gesund, entgegnen, er sei nur etwas weniger krank und auf der anderen Seite jemandem, der sagt, er sei krank, entgegnen, er sei nur etwas weniger gesund. Merkwürdigerweise entzieht dieses Paradox gerade der Begrifflichkeit ihren Sinn: Was kann jetzt als gesund bezeichnet werden? D.h. vor der Einführung des clinical-significance Begriffs steht die Auseinandersetzung um die Beschaffenheit des Heilungbegriffs, näher: die Auslegung dieser Begrifflichkeit in quantitativer Hinsicht. Zur weiteren Aufklärung trägt die folgende Distinktion bei (vgl. Hegel, 1832):

Qualitativer Unterschied: Ein Gegenstand ist von einem anderen verschieden, wenn mindestens ein Merkmal unterschiedlich ist. Wir sprechen, wenn dieses Merkmal zur Definition des Gegenstands wesentlich ist, von einem *anderen* Gegenstand. Ein Beispiel sei der Unterschied zwischen einem Apfel und einer Birne. Gesetzt

die Form der Birne sei der ausschlaggebende Unterschied[1] beider, wäre, falls diese Form einer Birne in die des Apfels geändert werden würde, die Birne ein Apfel.

Quantitativer Unterschied: Ändert sich eine Eigenschaft eines Gegenstandes in Richtung eines ‚mehr' oder ‚weniger', bleibt dieser derselbe (d.h. er verliert nicht seine Identität, wird nicht zu einem anderen). Beispiel ist etwa ein Stein, der, wenn er statt 5kg jetzt 10kg wiegt, doch noch ein Stein bleibt...

Maß: ...dass dieser Stein, bei weiterer Gewichtszunahme zu etwas anderen wird, etwa einem Fels oder Gebirge, deutet an, dass die quantitative Differenz zu einer qualitativen Differenz werden kann. Illustriert werden kann dies durch viele Fragestellungen der Psychophysik, etwa, ab welcher Tonhöhenänderung ein C wie H klingt. Hier interessiert vor allem die Grenze, ab welcher Maßzahl ein Gegenstand seine Qualität ändert. Bei dem ‚Maß' ist beides beteiligt, sowohl die quantitative als auch die qualitative Größe.

Der Hinweis auf die Psychophysik lenkt den Blick auf ein Forschungsgebiet, in dem ebenfalls die Formeln der ‚clinical significance' verwendet werden. Jedoch ist der Blickwinkel ein etwas anderer. In der Psychophysik geht es um Wahrnehmung: Ist der Gegenstand ein Apfel oder Birne, ist der Ton ein C oder H? Unmittelbare Ähnlichkeit hat diese Diskrimierungsaufgabe mit der diagnostischen: Ist eine Krankheit vorhanden? Es stellt sich also die Frage: Welcher Punkt des Kontinuums bezeichnet die Grenze zwischen Gesundheit und Krankheit? Die Antwort kann sehr unterschiedlich ausfallen:

- Bei der *Anorexia nervosa* ist das Gewichtskriterium das Hauptkriterium zur Diagnosestellung. Hier gibt es einen festgelegten Cut-Off zwischen gesund und krank, wobei dieser gewöhnlich durch medizinische Notwendigkeiten begründet wird. Die Umsetzung dieses Kriteriums erfolgt jedoch durch eine Konvention (BMI-Index, vgl. Oehlschlägel-Akiyoshi et al., 1999)

- Bei Abhängigkeitserkrankungen ist die Grenze bei der absoluten ‚Null' Punkt definiert. Nur bei fehlendem Alkoholkonsum spricht man von einem ‚trockenen Alkoholiker'.

Diese Beispiele illustrieren: es wird nicht die Gesundheit per se betrachtet, sondern nur ein isolierter *Aspekt* dieser. Diese Einschränkung ist natürlich verständlich: Ein Psychotherapeut behandelt die Essstörung, nicht den Beinbruch, also ist der Erfolg für diesen das Verschwinden des Syndroms ‚Essstörung' und allein dies erscheint im Forschungszusammenhang. Hinsichtlich des Gesundheitsbegriffs bedeutet dies, dass dieser etwas Zusammengesetztes ist. Dieser Umstand führt dann in der Forschung zu dem Problem multipler Erfolgskriterien: Erfolg ist vielleicht in einer besonderen Studie nicht nur das Verschwinden der Essstörung, sondern ebenso sehr der Abbau zwanghafter Charaktereigenschaften.

Diese oben angeführten Beispiele stellen hinsichtlich des Kriteriums Gesundheit / Krankheit jedoch eher eine Ausnahme dar, öfter liegt eine kontinuierliche Größe vor,

[1] Vgl. auch die alt-bewährte Definition der Definition ‚fit per genus proximum et differencia specialis'.

bei der die Grenze erst *gefunden* werden muss. Die Ausnahme besteht also im Vorhandensein einer *externen* Referenz zur Festlegung der Grenze. Welcher alternativer Mechanismus bietet sich hier an? Jacobson et al. (1984)bzw. Jacobson & Truax (1991) kompensieren das Fehlen eines externen Kriteriums durch den Blick auf die Verteilung des Merkmals in den beiden Gruppen, den Kranken und den Gesunden.

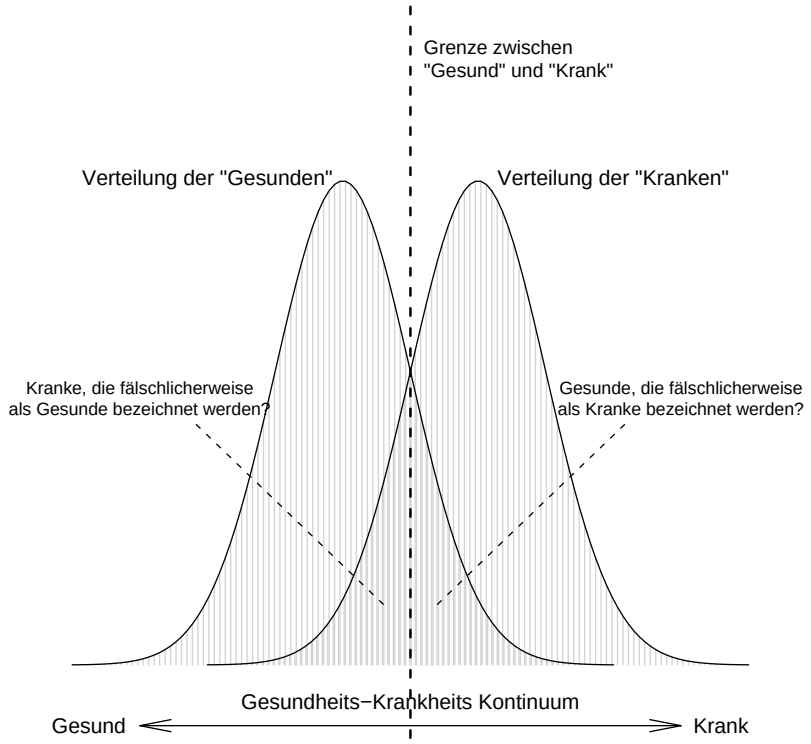

Abbildung 2.1: Darstellung der Umsetzung des ersten CS-Kriteriums: Differenz zweier Verteilungen.

Abbildung 2.1 versucht diesen Sachverhalt darzustellen. Auf der Abszisse ist hierbei das Gesundheits-Krankheitskontinuum dargestellt, auf der jede Person eindeutig lokalisiert ist. Die Höhe der Abbildung gibt die relative Anzahl der Personen, die an dem bestimmten Punkt auf der Abszisse lokalisiert sind, wieder. Dargestellt sind in der Abbildung zwei, von einander getrennte, eingipflige Verteilungen: die der Kranken und die der Gesunden. Weiterhin ist die aus der folgenden Formel (vgl. Formel 2.1 auf der nächsten Seite) sich ergebende Grenze zwischen Gesund und Krank eingezeichnet. Anhand dieser Grenze wird jetzt bestimmt, zu welcher Gruppe eine bestimmte Person gehört. Werden viele Personen derartig klassifiziert, ergibt sich eine absolute oder relative

Anzahl von Personen, die krank bzw. gesund sind.

$$CS = \frac{\mu_{x_{gesund}}\sigma_{x_{klinisch}} + \mu_{x_{klinisch}}\sigma_{x_{gesund}}}{\sigma_{x_{gesund}} + \sigma_{x_{klinisch}}} \quad \begin{cases} & \textit{Zugehörigkeit zur} \ldots \\ x \geq CS & \text{klinischen Population} \\ x < CS & \text{Population der Gesunden} \end{cases} \tag{2.1}$$

In der obigen Formel bezeichnet μ_x der zu erwartende Wert und σ_x die Variabilität der jeweiligen Subpopulation.

Diese Formel bildet in der Entwicklung des Begriffs einen weitreichenden Schritt, auch wenn viele implizite Annahmen noch verborgen sind. Wie auch eine Übersetzung eines Gedichts von einer Sprache in eine andere eigentlich unmöglich ist, da die Begriffe in den beiden Sprachen ganz andere Konnotationen annehmen, so erhält auch das bisher Erarbeitete weitere, bisher nicht vorhandene Spezifizierungen, allein durch die Übersetzung in die mathematische Sprachlichkeit (wobei die ersten beiden Punkte eher noch auf einer begrifflichen Ebene lokalisiert sind):

Zugrunde liegen *zwei* Gruppen: Die Tatsache, dass die Verteilungen zweier Gruppen bei der Umsetzung des Konzepts Voraussetzung ist, schafft einige Widersprüche:

- Wenn der Grenzpunkt (Cut-off) für jede Studie neu berechnet wird, kommt es zu dem Paradox, dass für jede Studie ein anderer Punkt als Grenze von gesund und krank gilt. Das heißt, dass Personen, die in Studie A als krank vielleicht bei der Studie B als gesund definiert werden. Handelt es sich hierbei um die gleichen Krankheitsbilder, so ist diese Variation der Cut-Offs vor allem auf zufälligen Schwankungen zurückzuführen, die sich über viele Studien ausgleichen sollten. Um die Vergleichbarkeit der Studien zu gewährleisten (wenn diese sich in der Definition von Krankheit unterschieden, wären sie nicht vergleichbar) schlagen Jacobson & Revensdorf (1988) vor, ähnlich der Gewinnung einer Normalstichprobe für Gesunde, eine Normalstichprobe für Kranke zu erheben und einen für alle Studien verbindlichen Cut-Off zu bestimmen. Dieser Vorschlag wurde jedoch nach Wissen des Autors bisher nicht umgesetzt.
- Dieses Problem löst jedoch einen anderen verwandten Einwand nicht: Wenn die gleiche Dimension betrachtet wird, etwa Depressivität und ein Cut-Off für verschiedene Krankheitsbilder bestimmt wird, also etwa für die Dysthymia und die endogene Depression, ergeben sich auch hier unterschiedliche Cut-Offs und mithin unterschiedliche Normalitätsgrenzen für verschiedene Krankheitsbilder (Kordy, 1997). Es ergäbe sich also wieder ein ähnliches Problem: so könnte jemand in der einen Studie noch als gesund und in einer anderen Studie als krank bezeichnet werden.

Form der Verteilung: Die Annahme einer Normalverteilung beider Gruppen ist unplausibel. Dies wurde von den Autoren auch selbst gesehen (Jacobson & Revensdorf, 1988). Diese Tatsache stellt jedoch kein besonders gravierendes Problem dar,

sind doch ausreichend nichtparametrische Alternativen vorhanden, so etwa die Receiver-Operating-Curves (Bregg, 1991; Mossman & Somoza, 1989).

Wert von Falschklassifikationen? In der oben genannten Abbildung ist implizit folgende Bewertung enthalten: Wenn falsch klassifiziert wird, dann so, dass genauso viele Gesunde fälschlicherweise als krank und Kranke fälschlicherweise als gesund bezeichnet werden. Wäre dies nicht so, wäre das Ergebnis hinsichtlich des Erfolgs entweder zu optimistisch oder zu konservativ[1]. D. h. implizit ist in der Formel eine *Wertschätzung* verborgen. Im diagnostischen Prozess ist der Fehler, etwas nicht zu finden oder fälschlicherweise etwas zu diagnostizieren nicht immer gleich gravierend: Als Krankheitsscreening sollten keine möglichen Krankheiten übersehen werden, vor einer schweren Operation dagegen sollte die Diagnose absolut gesichert sein.

Kranke in der Population der Gesunden? Dieser Einwand hängt mit dem obigen zusammen, beleuchtet das Problem jedoch von einer anderen Seite. Die Population der Gesunden ist scheinbar klar definiert. Wird diese jedoch erhoben, sind natürlich einige nicht behandelte Kranke in dieser enthalten. Das bedeutet, dass in der Fläche, die in Abbildung 2.1 mit ‚Gesunde, die fälschlicherweise als Kranke bezeichnet wurden' beschrieben ist, sich ein (unbekannter) Anteil an Kranken verbirgt, die richtigerweise als ‚krank' etikettiert werden. Jacobson & Revensdorf (1988) schlagen daher vor, nur ‚reine' Stichproben zu erheben, also eine Stichprobe der Erkrankten ohne Gesunde und vice versa.

Die hier ausgeführten Punkte stellen teilweise eine massive Kritik des CS-Ansatzes dar. Besonders der erste Punkt, der zeigt, dass Normalität in diesem Ansatz ein relativer Begriff ist (relativ zu der untersuchten Krankheit), ist schwer aufzuräumen. Eine alternative Möglichkeit besteht darin, analog zur üblichen Verfahrensweise in der psychologischen Testung von nur einer Stichprobe auszugehen und die Grenze etwa durch die Angabe durch Perzentile zu bestimmen. Jacobson & Revensdorf (1988) haben dies als Ausweg angegeben, wenn nicht zwei Stichproben vorliegen; Kordy (1997) schlägt vor, dies aufgrund der oben beschrieben Problematik durchgängig in einer derartigen Weise zu handhaben. Hier stellt sich das den oben genannten Problemen verwandte Problem, dass auch eine Perzentilangabe willkürlich ist: Soll 1% oder 5% der Stichprobe als krank klassifiziert werden? Diese Konzeption wird in Abschnitt 2.2.3 auf der nächsten Seite nochmals ausführlich dargestellt.

2.2.2 Zusammenfassende Betrachtung

Bevor dieser alternative Ansatz vorgestellt wird, sollte die bisherige Argumentation zusammengefasst werden. Ausgegangen wird von dem *Begriff der Gesundheit* mit

[1] Dies sei hier insbesondere wegen eines anderen Kontextes, in dem diese Abbildung gesehen werden muss, erwähnt. Im Kontext des statistischen Schließens (Null-Hypothesen Tests) repräsentiert die linke Verteilung die Null-Hypothese, die rechte die Alternativhypothese. Wie bekannt, wird der Cut-Off hierbei nicht symmetrisch, sondern zugunsten des so genannten alpha-Fehlers gebildet (für Details siehe Abschnitt 3).

seinen Eigenschaftswörtern gesund und krank. Es zeigt sich, dass es auf begrifflicher Ebene gute Gründe gibt, diese Dimensionen als Kontinuum anzusehen. Also wird dieser Begriff in eine *kontinuierliche Dimension* übersetzt, die an sich eine *Grenze* hat, die die Pole gesund und krank trennt. D.h. es handelt sich um ein Maß.. Durch die Übersetzung in eine mathematische Formelsprache werden die beiden Abschnitte des Kontinuums als *Verteilungen* gefasst und die Grenze beider ist nunmehr die Grenze der beiden Verteilungen. Dies bedeutet: es muss schon vorher bekannt sein, wer zu den beiden Gruppen gehört, diese werden beobachtet und es wird das Kriterium bestimmt, das in Zukunft bestimmen soll, wer zu der Gesunden- und Kranken-Gruppe gehört. Dies kann auch als *Münchhausenprinzip* (sich selbst an dem Schopf aus der Grube ziehen, engl.: ‚Bootstrap') bezeichnet werden: Man weiß, was als gesund und krank zu gelten hat und benützt diese Information, um in einem anderen Medium herauszufinden, was gesund und krank ist. Es zeigt sich, dass diese Art der Konzeptualisierung von impliziten Werten (Wert der Falschklassifikation) und technischen Problemen begleitet ist.

Als Resultat wird in der konkreten Anwendung also jedes Individuum klassifiziert. Ergebnis der Klassifikation ist eine absolute Anzahl (200 von 300 sind ...) oder eine relative Anzahl ($2/3$ der ...).

Hinsichtlich des *Gesundheitsbegriffs* ergibt sich folgende, verdichtete Zusammenfassung: Durch diese Konzeptualisierung wird der Begriff der Gesundheit zu etwas Zusammengesetztem. Die Teilaspekte dieses Zusammengesetzten sind hierbei kontinuierliche Größen, welche mit einer Grenze versehen sind. Die Krankheitsgrenze ist bestimmt durch externe Kriterien (besonders bei medizinischen Erkrankungen), bei psychologischen Erkrankungen durch Verteilungen (Normalitätsbegriff). Der menschliche *Körper ist also eine Vielheit von Verteilungen*, wobei nicht klar ist, wie sich diese Vielheit zum abstrakten Begriff der Gesundheit zusammensetzt (additiv, oder per ‚und' verknüpft?).

2.2.3 Eine alternative Definition der klinischen Signifikanz

Gewöhnlich wird die Abweichung von der Normalität nicht, wie oben dargestellt, durch zwei Verteilungen bestimmt. Üblich ist die Angabe des Cut-Offs der Gesamtgruppe, so etwa das 5%-Perzentile. Auch hier kann zur Bestimmung der Grenze das oben erwähnte *Münchhausenprinzip* angewandt werden. So könnte vielleicht durch epidemiologische Untersuchungen bekannt sein, dass ein bestimmter Anteil einer Population erkrankt ist, die Grenze zwischen den Erkrankten und Gesunden auf der jeweiligen Dimension könnte also im folgenden als Grenze fungieren, was in Zukunft gesund und krank sein soll. Veranschaulicht werden soll dies noch einmal durch eine Abbildung (2.2). Bei dieser Abbildung ist analog zu der Abbildung 2.1 auf Seite 9 auf der Abszisse das Gesundheits-Krankheitskontinuum abgetragen. Wie auch bei der Abbildung 2.1 entspricht die Höhe der Kurve dem relativen Anteil der Personen, die auf einem bestimmten Punkt auf der Abszisse lokalisiert sind.

Auch diese Umsetzung basiert auf dem oben herauskristallisierten (vgl. 2.2.2) Mechanismus. Es gibt jedoch einen wichtigen Unterschied: Es werden nicht mehr *zwei*

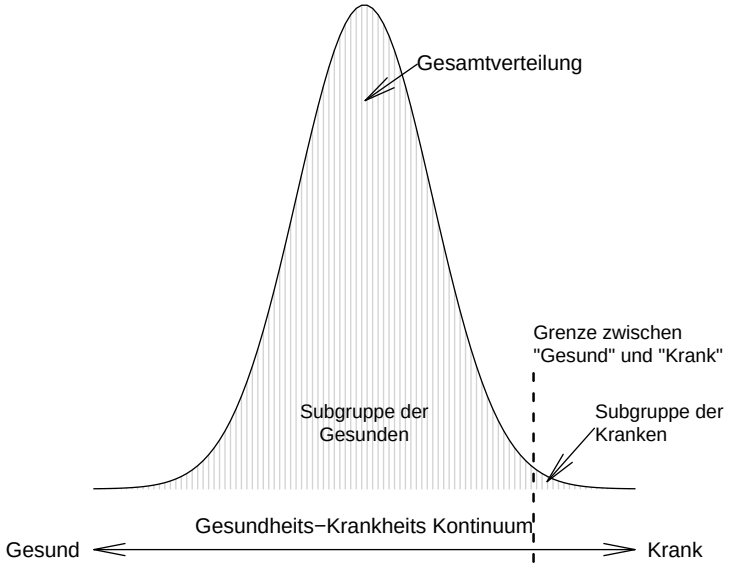

Abbildung 2.2: Darstellung der Umsetzung des ersten CS-Kriteriums: Perzentile einer Verteilung..

Verteilungen betrachtet. Hier ergibt sich in Folge ein sehr praktisches Problem: Da Messungen an den Rändern der Verteilung sehr unsicher sind (eine Prävalenz von 1% bedeutet bei einer Stichprobe von 100 Personen und einem 95% Vertrauensintervall von 0.02 bis 5.4, bei 1000 Personen von 0.48 bis 1.83, bei 10000 Personen von 0.8 bis 1.2) sind große Stichproben zur Normierung notwendig.

2.3 Der zweite Teil des Indizes: das reliable change Kriterium (RC)

Im Folgenden wird das reliable-change Kriterium (RC) eingeführt und gleichzeitig die Einbettung dieses Kriteriums in den gesamt CS-Begriff dargestellt. Am Ende des Abschnittes stehen methodische Fragen im Vordergrund.

2.3.1 Begriffliche Analyse und Ableitung des Indizes

Ausgangspunkt ist wie oben der Begriff der Heilung, jedoch wurde im vorigen Abschnitt der Akzent auf die Beurteilung des Status gelegt. Bei dem Reliable Change Kriterium (RC) wird auf den Wechsel beider Zustände, also auf *Veränderung* gezielt. Der Begriff der Veränderung impliziert den der Differenz. Genau genommen handelt es sich um eine Differenz, welche an einem Gegenstand stattfindet: Gegenstand A ist mit Eigenschaft B versehen, welche später durch C ersetzt wird, oder in der obigen Formulierung des Maßes (siehe die Unterteilung in Abschnitt 2.2.1 auf Seite 7): Ein Gegenstand verändert sich so lange, bis er ein anderer wird.

Das Wort ‚später' verweist hier schon auf ein Geschehen in der Zeit. Es ist hier zu bemerken, dass andere Veränderungen, etwa logische, durchaus für den Leser in der Zeit stattfinden können, an sich jedoch außerhalb dieser liegen (beispielsweise logische Zusammenhänge). Im folgenden werden jedoch ausschließlich empirische Veränderungen betrachtet.

Näher betrachtet handelt es sich um *wahre* Veränderungen. Warum ‚wahr'? Was lässt mich daran zweifeln, dass der Patient X auf die Frage ‚Wie fühlen Sie sich' einmal ‚nicht so gut' dann wieder ‚gut' ankreuzt? und jeweils seinen Zustand korrekt beschreibt?

Begründet wird diese in Fragestellung des Ankreuzverhaltens durch die Messtheorie: Das Instrument, mit dem wir den Gegenstand betrachten, wird häufig als trübe Linse vorgestellt. Der Zugang zur Realität ist also durch ein ungenaues Medium getrübt. Diese Erklärung erfährt ihre Plausibilität aus einem alltäglichen Verständnis des Messens, im Kontext der Messtheorie ist dieser jedoch nicht zutreffend. Der nächste Abschnitt mag aufzeigen, was der wahre Wert sowie der Fehler in diesem Kontext bedeutet.

Exkurs: Die verschiedenen wahren Werte

Leider sind die Arbeiten, die sich mit der Begründung des Begriffs des Messfehlers bzw. des wahren Wertes beschäftigen, in Vergessenheit geraten. An dieser Stelle wird im wesentlichen der Gedankengang des Kapitels ‚The many concepts of true scors' aus dem Klassiker der Messtheorie Lord & Novick (1968) wiederholt:

Der ‚platonische' wahre Wert: Es wird angenommen, dass außerhalb des Experimentes ein wahrer Wert existiert. Beispiel von Sutcliffe (1965) (der selber dieses Konzept nicht vertrat) ist die Bestimmung des Geschlechts bei jungen Hühnern. Ob es sich um eine Henne oder Hahn handelt, steht unabhängig fest, die mögliche Fehlerquelle liegt allein beim Beurteiler. Besonders die Faktorenanalyse wurde anhand dieser Konzeption begründet. Wegen der Wagheit psychologischer Theorien ist sie jedoch nicht in der Psychologie realisierbar. Auch würde diese zu paradoxen Situationen führen, würde sie mit der klassischen Testtheorie verbunden.

Der wahre Wert als Grenzwert: das Experiment wird unendlich oft wiederholt. Der Grenzwert (‚limiting value') des Durchschnittswerts konvergiert nun zu einer

Konstanten, also dem wahren Wert. Mises (1919) Konzeption der Wahrscheinlichkeit konnte sich jedoch aufgrund von technischen Schwierigkeiten nicht durchsetzen.

Operationale Definition des wahren Wertes ähnelt sehr der obigen Konzeption, verzichtet aber auf den Unendlichkeitsbegriff[1]. Illustriert werden kann dies anhand des Standardbeispiels von Lazarsfeld (1959): Mr. Brown wird wiederholt gefragt, was er für eine Meinung bzgl. der Vereinten Nationen hat. Zwischen den wiederholten Fragen unterzieht er sich einer Gehirnwäsche, so dass er vergisst, was er geantwortet hat und praktisch am gleichen Punkt wieder anfängt. Da er sich seiner Sache nicht sicher ist, gibt er manchmal positive, manchmal negative Einschätzungen. Die relative Anzahl der positiven Antworten ist hier also der wahre Wert des Mr. Brown[2]. Wird diese Prozedur jedoch angewandt, wenn Mr. Brown angetrunken ist, mag er vielleicht andere Antworten geben, da einige seiner Hemmungen wegfallen. D. h. der wahre Wert bezieht sich ebenso auf einen Zustand der Person. Diese State-Trait Theorie wurde dann später von Steyer, Ferring, & Schmitt (1992), Steyer & Schmitt (1998) explizit formuliert.

Wie ersichtlich, ist der intuitiv gemeinte Begriff des ‚wahren Wertes' nicht der in der klassischen Testtheorie verwendete. Es handelt sich hier um eine ‚operationale' Definition. Interessant ist hierbei, dass das Paradigma dieser Operationalität im Imagiären liegt. So ist das radikale ‚Kopf waschen' im obigen Beispiel zwar gut vorstellbar, entspricht jedoch kaum einer denkbaren Realität.

Was bedeutet nun der Begriff des ‚Fehlers'? Fehler besagt: Nicht kontrollierbare experimentelle Faktoren. Hier offenbart sich die auch relative Natur des ‚wahren Wertes': relativ nämlich zur experimentellen Situation und den dort kontrollierten Bedingungen.

2.3.2 Zusammenfassende Betrachtung und Einbettung in den gesamten CS-Begriff

Der RC-Begriff ist ein Urteil, welches eine beobachtete Veränderung dahingegen beurteilt, ob sie als zuverlässig gelten kann, d.h. dass sie eine wirkliche Änderung ist. Mithin resultieren zwei Ausssagen: (1) Es fand sich eine Verschlechterung/Verbesserung (2) es fand sich keine Änderung. D.h. bei dem Urteil auf CSK wird in der Zeit eine Änderung des Zustandes gefordert, während bei dem Urteil auf RC der Ausmaß der Änderung beurteilt wird.

[1] Genau genommen ist die operationale Definition in die ‚von-Mises-Definition' überführbar: Vgl. das Konzept des unendlich langen Tests (Lord & Novick, 1968).

[2] Besser darstellen lässt sich dieser Sachverhalt auf der Ebene der Imagination: Mr. Brown wird zu den Vereinten Nationen befragt, die Zeit wird mit einer Zeitmaschine zurückgedreht und er wird wieder befragt. Dieses Beispiel verzichtet auf eine explizite Manipulation des Zustandes von Mr. Brown (Gehirnwäsche). Es ist jedoch anzunehmen, dass diese Modifikation schwerlich die Zustimmung des Autors gefunden hätte, verweist sie doch explizit auf den imaginären Charakter des Beispiels: In der realen Welt ist eine solche Situation nicht anzutreffen.

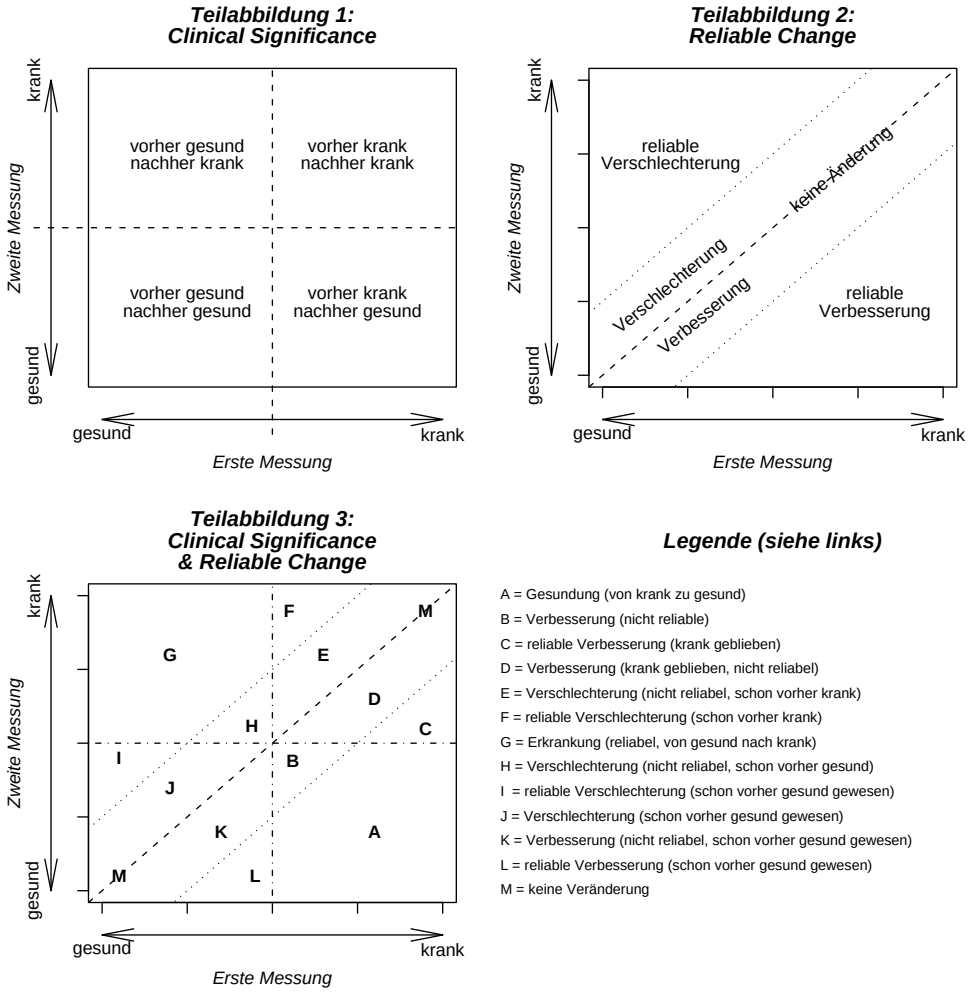

Abbildung 2.3: Integration der beiden Kriterien (CSK & RC) des CS-Ansatzes.

Abb. 2.3 zeigt die einzelnen Aspekte dieser beiden Urteile. In dieser Abbildung sind drei Teilabbildungen integriert. Bei allen drei Abbildungen handelt es sich um so genannte Scatterplots, bei dem die Abszisse den ersten Messzeitpunkt wiedergibt, die Ordinate den zweiten Messzeitpunkt. Links oben ist eine vereinfachte Darstellung des CSK-Kriteriums zu sehen, rechts oben eine Abbildung des RC-Kriteriums und links unten eine Integration der beiden Ansätze. Bei allen drei Teilabbildungen handelt es sich um stilisierte Scatterplots, bei denen die Abszisse den ersten Messzeitpunkt realisiert, die Ordinate den zweiten Messzeitpunkt. Was zeigt nun die in der Literatur durchweg verwendete und dadurch fast den Status des Paradigmatischen erreichen-

de Illustration[1] des CS Begriffs? In diesem werden zwei kontinuierliche Größen miteinander in Beziehung gesetzt. Jede Person ist dabei sowohl durch die eine als auch durch die andere eindeutig bezeichnet. Die beiden Achsen unterscheiden sich vor allem durch die zeitliche Differenz zwischen ihnen: so könnten diese u.U. Messzeitpunkte wie Aufnahme-Entlassung, Aufnahme-Katamnese, Entlassung-Katamnese repräsentieren. Diese Achsen sind jetzt wiederum Repräsentanten des Gesundheits-Krankheits-Kontinuums, und wiederum lässt sich auch in diese Bildlichkeit die Grenze zwischen Gesundheit-Krankheit einzeichnen. Wird hier die in der Erläuterung des ersten Teils des ‚clinical significance'-Begriffs eingeführte Grenze (Cut-Off) zwischen Krank und Gesund eingetragen, ergibt sich eine Vierfelder-Tafel. Visualisiert ist dies in Teilabbildung 1 der Abbildung 2.3. Verknüpft mit der in der Abbildung eingeführten zeitlichen Dimension ergeben sich die Begriffe, die eine Änderung anzeigen: Vorher krank - nachher gesund (Heilung), vorher gesund, nachher krank (Erkrankung) sowie die beiden Begriffe, die eine Stagnation bezeichnen: Vorher und nachher gesund und vorher und nachher krank.

Wie ist nun der *Unterschied* beider Messzeitpunkte visualisiert? Als Abweichung von der Diagonale (siehe Teilabbildung 2 der Abbildung 2.3). Liegt ein Wert auf dieser, bedeutet dies, dass er zu beiden Messzeitpunkten den gleichen Wert erreicht. Nun ist jedoch, wie gesagt, jede Messung mit einem Fehler behaftet und mithin auch die Differenz beider. Dieser Fehler ist das in der Abbildung eingezeichnete ‚Band' um die Diagonale. Liegt jetzt ein Messzeitpunkt in dieser Region, hat man zwar eine Veränderung festgestellt, jedoch ist nicht klar, ob diese nicht durch den Messfehler bedingt ist.

Führt man beide Teilabbildungen und mithin beide Teilkonzepte zusammen, ergeben sich verschiedene ‚Schnittpunkte':

Gesundung (Feld: A): Nach dem Konzept der clinical significance liegt hier eine *Heilung*, oder Gesundung vor: Vorher krank, nachher gesund und die Veränderung liegt auch über dem Messfehler.

Erkrankung (Feld: G): Dieser Fall ist im Kontext einer psychotherapeutischen Intervention als der schlechteste Fall zu bezeichnen: Vorher gesund, nachher krank und die Differenz ist als zuverlässig anzusehen.

Verbesserung/Verschlechterung, jedoch ohne RC (Feld: H, B): Die beiden obigen Fälle, also die Gesundung und die Erkrankung sind natürlich auch ohne das At-

[1] Visualisierungen sind immer auch Instrumente der Interpretation. So wie sie *etwas* zeigen, verbergen sie *anderes*. Das kritische Bewusstsein hinsichtlich der Visualisierung ist jedoch relativ jung, besonders die Arbeiten von Cleveland (1993, 1994) haben dazu beigetragen, diese in ihren Möglichkeiten und Grenzen wahrzunehmen. Ausschlaggebend für dieses meist unbewusste Spiel (wäre dies dem Forscher bekannt, handelte es sich um eine bewusste Irreführung.) des Zeigens und Verbergens ist die auch durch Sehgewohnheiten geschulte menschliche Wahrnehmung. Bekanntestes Beispiel sind hier wohl die bekannten Tortengrafiken, die versprechen quantitative Verhältnisse zu zeigen (etwa Prozente), dies jedoch nicht einlösen können, einfach deshalb, weil der Mensch schlecht Flächen vergleichen kann (Cleveland, 1994). D.h. die verbreitete Darstellungspraxis ignoriert basale Mechanismen menschlichen Sehens.

tribut ‚reliable' denkbar: D. h. es gab eine Änderung, jedoch lag diese nicht außerhalb des Bereichs der Messfehlerunsicherheit.

Variationen im Bereich des Gesunden (Feld: I, L, J, K):] Hier kann es zu einer reliablen Verbesserung (L)oder Verschlechterung (I) kommen oder es kommt ohne das Attribut ‚reliable' zu einer Verschlechterung (J) oder Besserung (K).

Variation im Bereich des Kranken (Feld: F, E, D, C)] Auch hier kann es, spiegelbildlich zum obigen Punkt, zu reliabler Verbesserung (C), Verbesserung (D), reliabler Verschlechterung (F) oder Verschlechterung (E) kommen, wohlgemerkt im Bereich des Kranken.

keine Änderung (Feld: M) Und natürlich kann auch der Zustand gleich bleiben, diese Möglichkeit wird durch die Diagonale repräsentiert.

Es ergibt sich also, alleine aus der Kombination der beiden Teile des clinical significance Ansatzes ein neuer, vorher nicht definierter, Bedeutungsraum: Den Möglichkeiten der Veränderung. Dabei muss allerdings beachtet werden, dass ein Großteil der oben erwähnten Kombinationen eigentlich entfallen. Dies betrifft alle innerhalb des diagonalen Streifens liegenden Bestimmungen der Messfehlerunsicherheit. Genau genommen werden acht der genannten Möglichkeiten (M, J, K, H, B, E, D, M) mit ‚keiner Veränderung' zusammengefasst. Dieses Feld der Differenzen findet sich jedoch nicht in dem zusammenfassenden Index wieder (siehe Formel 1.1 auf Seite 2). Durch die Und-Verknüpfung sind nur noch zwei Fälle von Interesse:

Erfolg (Feld: A): Der Patient war vor der Behandlung krank, die Veränderung war positiv und reliable.

Misserfolg (alle außer: Feld A): Der Patient war entweder schon gesund (G, H, I, J, K, L), war krank und hat sich nicht geändert (D, E, F) oder hat sich zwar geändert, erreicht jedoch nicht den gesunden Bereich (C)

Es handelt sich also bei diesem Schritt um eine weitere Verdichtung von Ergebnissen. Im Verhältnis zur vorherigen Fülle an Differenzierungen ist dies jedoch auch ein Verlust an Information.

2.3.3 Methodische Aspekte des RC-Kriteriums

Im Gegensatz zur relativen methodischen Eindeutigkeit des CSK-Kriteriums, liegen hier sehr viele divergente Konzeptualisierungen des Begriffs des RC vor. Tabelle 2.3 gibt einen Überblick zu den in der Literatur zu findenden Umsetzungen.

Diese verschiedenen Umsetzungen versuchen verschiedenste Korrekturen einzuführen, um Phänomene wie die ‚regression to the mean' zu bewältigen. Wie ersichtlich, verkomplizierten sich, ausgehend vom Originalansatz, die Formeln dadurch. Diese wachsende Komplizität kann hier als Reflex der wachsenden Berücksichtigung von verschiedenen Phänomenen verstanden werden. Auch ist ein fundamentaler Unterschied zu

Die verschiedenen "reliable change" Definitionen

Formel	**Quelle**	**Beschreibung**
$RC = (X_1 - X_2)/S_E$	$S_E = S_1\sqrt{1 - r_{XX}}$	Jacobson, Folette, & Revensdorf (1984)
$RC = (X_1 - X_2)/S_D$	$S_D = \sqrt{2S_E^2}$	Christensen & Mendoza (1986)
$RC = (T_1 - X_2)/S_D$	$T_1 = r_{xx}(X_1 - \overline{X}_1)$	Nunnally & Kotsch (1983), Hsu (1989)
$RC = (T_1 - \overline{X}_1) - (X_2 - \overline{X}_2)/S_{Pred}$	$S_{Pred} = s_1\sqrt{1 - r_{xx}^2}$	Speer (1992)
$RC = (T_1 - T_2)/S_E$	$RC = T_1 - T_2 = r_{dd}(X_1 - X_2) + (1 - r_{dd})(\overline{X}_1 - \overline{X}_2)$ und $S_E = \sqrt{s_{E_1}^2 + s_{E_2}^2}$ und $s_{E_1} = S_1\sqrt{1 - r_{XX_1}}$ und $s_{E_2} = S_2\sqrt{1 - r_{XX_2}}$ und $r_{DD} = \frac{s_1^2 r_{XX_1} + s_2^2 r_{XX_2} - 2s_1 s_2 r_{12}}{s_1^2 + s_2^2 - 2s_1 s_2 r_{12}}$	Hageman & Arrindell (1993)
$RC = (T_1 - T_2)/S_{PredDiff}$	$S_{PredDiff} = S_{Ddif}\sqrt{r_{dd}(1 - r_{dd})}$ und $S_{Ddif} = \sqrt{s_1^2 + s_2^2 - 2s_{12}^2}$	Zegers & Hafkenscheid (1994)

wobei:

RC	Reliable change Index	X_n	Gemessener Wert zu Zeitpunkt n
S_n	Standardabweichung zum Zeitpunkt n	r_{xx}	Retestkorrelation
S_D	Standardfehler der Differenzen	T_n	Geschätzter wahrer Wert zu Zeitpunkt n
r_{dd}	Reliabilität der Differenzen	S_{Pred}	Standardfehler der Prädiktion

Tabelle 2.3: Überblick zu den verschiedenen RC-Konzeptionen

der Originalformel (siehe 2.2) zu konstatieren: In den letzten vier Formeln wird versucht, den wahren Wert T zu einem (Nunnally & Kotsch, 1983; Hsu, 1989; Speer, 1992) oder zu beiden Zeitpunkten zu schätzen (Hageman & Arrindell, 1993; Zegers & Hafkenscheid, 1994), d.h., es wird nicht mit den beobachteten, sondern mit den korrigierten Werten operiert.

> „Different RC's may lead to different conclusions concerning the effect of an intervention on a given person, a cause of much confusion" (Maassen, 2000b)

Maassen (2000b) stellt in dem obigen Zitat dar, was viele vergleichende Arbeiten zeigen (so etwa Nunnally & Kotsch, 1983): Unterschiedliche RC-Indizes ergeben unterschiedliche Ergebnisse. Der Mangel derartiger Vergleiche ist offensichtlich: Es fehlt das *Kriterium* zur Entscheidung, welches die *richtige* Formel ist. In einer ganzen Reihe von kritischen Arbeiten konnte Maassen (2000a, 2000b, 1998) zeigen, dass die Arbeiten sowohl ihre eigene Zielsetzung, den Messfehler zu korrigieren, nicht genügen und vielmehr irreführende und verfälschende (biased) Ergebnisse produzieren. Es handelt sich bei den kritisierten Arbeiten um Versuche, zwei verschiedene Ansätze zu kombinieren, nämlich die Schätzung der wahren Werte aus den beobachteten und die Beurteilung, ob die beobachteten Werte außerhalb des durch die Messfehlerungenauigkeit bedingten Bereiches liegen. Eine Reduktion dieser Ansätze auf die sie begründende Prinzipien zeigte hierbei, dass nur zwei unterschiedliche Ansätze zu finden sind. Die Vielzahl der Formeln gründet alleine auf Variation dieser beiden Prinzipien.

- Bei *dem klassischen Ansatz*, wird ein Urteil gefällt, ob eine Messwertdifferenz außerhalb des durch den durch Messfehler bedingten Bereich liegt, d.h. die Messwerte werden verworfen.

- Dagegen werden bei dem *weighted* Ansatz die individuellen Messwerte je nach Ausmaß des Messfehlers korrigiert.

Der klassische wie der ‚reliable weighted' Ansatz sind asymptotisch gleich, d.h. bei der Anwendung großer Stichproben oder sehr häufiger Wiederholung kommt es zu den gleichen Ergebnissen. So kommt Maassen (2000b) zu folgendem Schluss:

> „All in all, in our view there are strong arguments for preferring the classic approach to the estimate interval method. This method has been undeservedly regarded inferior by the authors who recently proposed new indices in the clinical psychology literature." Maassen (2000b, S. 631)

Jedoch ergeben sich in bestimmten Situationen Unterschiede; bevor diese Situationen beschrieben und eine abschließende Bewertung dieser Schlussfolgerung vorgenommen wird, seien diese beiden Ansätze kurz umrissen.

Die Originalformel: kritische Differenz

Die Orginalformel lautete folgendermaßen:

$$(D_{krit})_{0.05} = 1.96\sigma_x\sqrt{2(1-r_{xx})} \begin{cases} x \geq D_{krit} & \text{reliable Verbesserung} \\ |x| < D_{krit} & \text{keine Veränderung} \\ x \leq (-D_{krit}) & \text{reliable Verschlechterung} \end{cases} \tag{2.2}$$

wobei r_{xx} die Reliabilität eines Messinstruments und 1.96 die zweiseitigen Irrtumswahrscheinlichkeit nach der Standardnormalverteilung bezeichnen. Voraussetzungen sind die Annahmen der Normalverteilung, Gleichheit der Streuungen $s_X = s_{X1} = s_{x_2}$ und der Reliabilitäten $r_X = r_{x_1} = r_{x_2}$ des Messinstruments[1].

Grundannahme ist (vgl. Gleichung 2.3), dass sich die beobachtete Differenz D_i eines Individuums i aus der wahren Differenz Δ_i und einem Messfehler zusammensetzt ist:

$$D_i = Y_i - X_i = \Delta_i + E_{D_i} \tag{2.3}$$

In der Regel wird angenommen, dass der Fehler normalverteilt ist sowie eine Standardabweichung von σ_{E_D} hat (Lord & Novick, 1968, S. 159). Unter der Null-Hypothese, dass das Treatment keinen Effekt hat besteht eine Standard-Normalverteilung. Wenn dieser Wert nun einen bestimmten Wert, etwa um 1.96 überschreitet, kann die **Null-Hypothese: ‚Die wahre Veränderung eines Individuums ist 0'** abgelehnt werden. Wird nun der Fehlerterm σ_{E_D} spezifiziert, ergibt sich wohl die bekannteste Formel RC Jacobson & Truax (1991), Jacobson et al. (1984) (siehe 2.2).

Der ‚reliable weighted'-Ansatz

Der grundlegende Unterschied zum klassischen Ansatz ist, dass keine individuellen Messwerte abgelehnt werden. Stattdessen werden diese in Richtung des Gruppenunterschieds korrigiert.

$$\rho_{DD}D_i - (1-\rho_{DD})\overline{D} \pm 1.96\sigma_{\Delta.D} \tag{2.4}$$

wobei

$$\sigma_{\Delta.D} = \sigma_\Delta(1-\rho_{DD})^{1/2} = \sigma_D(\rho_{DD})^{1/2}(1-\rho_{DD})^{1/2} \tag{2.5}$$

oder etwas handlicher:

$$\rho_{DD}D_i - (1-\rho_{DD})\overline{D} \pm 1.96\sigma_D\sqrt{2-\rho_{DD}^2} \tag{2.6}$$

So in der obigen Formel 2.4. Die individuell beobachtete Differenz D_i wird mit der Reliabilität des Instruments r_{DD} gewichtet $r_{DD}D_i$, d.h. wenn die Reliabilität des Instruments im Extremfall Null wäre, wird dieser Ausdruck auch Null. Was passiert dann?

[1] Kritisch einzuschätzen ist, wie schon beim clinical significance Kriterium, die Normalverteilungsannahme: Bei klinischen Skalen sind bei Nichterkrankten oft nur wenige Veränderungen zu erwarten, d. h. es wird eine schiefe Verteilung zu erwarten sein.

Der zweite Teil der Formel $(1 - \rho_{DD})\overline{D}$ verkürzt sich zu $(1 - 0)\overline{D} = \overline{D}$, so dass der individuelle ‚beobachtete Wert ‚verschwindet' und nur die beobachtete Differenz in der Gruppe beachtet wird. Dieser oben beschriebene Fall ist jedoch natürlich nur ein Grenzfall. *Es wird bei diesem Ansatz der individuelle Wert desto mehr dem Gruppenwert angeglichen, je geringer die Realiabilität des Wertes ist.*

2.4 Zusammenfassende Darstellung und abgeleitete Definitionen

Je nach Ebene der Betrachtung ergibt sich ein ganzes Bündel an möglichen Definitionen des ‚clinical significance' Konzeptes (CS), mit seinen beiden untergeordneten Konzepten clinical significance (CSK) und reliable change (RC). Will man den Bedeutungsraum des Konzepts und seiner beiden Teile umreißen, sind wohl die folgenden Begrifflichkeiten relevant:

clinical significance: Die direkte Übersetzung lautet: klinische Bedeutsamkeit. Mit dem Teilkonzept vernüpfbare Begriffe wären: ‚Gesundung', ‚Gruppenzugehörigkeit', ‚Anzahl der Gruppenwechsel'. Wird hingegen stärker auf die methodische Ebene fokussiert, ist vor allem an den Begriff der ‚Normalität' zu denken.

reliable change: Die direkte Übersetzung lautet: zuverlässige Änderung. Hier wären verknüpfbare Begriffe: ‚wirkliche (vs. scheinbare?) Veränderung', ‚bedeutsame Veränderung', ‚zuverlässige Änderung', ‚Anzahl der Änderungen'. Auf der methodischen Ebene ist dieser Bedeutungsraum jedoch auch den Begriff des Messfehlers eingeschränkt.

Das ganze Konzept: Es fällt schwer hier etwas anderes als ‚klinische Signifikanz' vorzuschlagen. Mögliche Begrifflichkeiten wären: ‚klinisch begründetes Erfolgsmaß', ‚klinischer Erfolg', o.ä.

Die Einführung neuer Begrifflichkeiten ist immer problematisch, wird diese kaum von allen Adressaten auch nur zur Kenntnis genommen. Aus diesem Grund wird an der *direkten* Übersetzung der englischen Begrifflichkeit festgehalten: ‚klinische Signifikanz' (CS) für das gesamte Konzept und den ersten Teil des Indizes (CSK) und ‚reliable change' (RC) für den zweiten Teil des Indizes.

Wie sieht nun die Definition der „klinischen Signifikanz" aus?

> Unter dem Begriff CS werden diejenigen Erfolgsmaße bezeichnet, bei denen die folgenden beiden Kriterien beide erfüllt sein müssen. (1) Es muss ein Wechsel von einem krankheitswertigen zu einem nicht mehr krankheitswertigen psychischen Zustand vorliegen (CSK). (2) Es muss eine wirkliche Veränderung vorliegen (RC).

Bei der obigen Definition stehen die beiden Teilkriterien CSK und RC gleichwertig nebeneinander. Der zweite Teil der Definition, der den RC-Teil beschreibt, ist bewusst

vage gehalten: Was eine wirkliche Änderung sein soll, wird sich erst noch ergeben. Im Kontext des klassischen RC-Begriffs würde alleine der Messfehler Berücksichtigung finden.

3 Exkurs: Abkehr von der Wahrscheinlichkeit als Begründung im Forschungsgeschehen

Wichtigstes Mittel wissenschaftlichen Fortschritts in der Psychologie ist zunächst und zumeist der Signifikanztest. Dieser ist Moment im Fortschreiten der Psychologie, dem Hinzukommen neuen Wissens in diesem Gegenstandsbereich. In verschiedenen Forschungszusammenhängen hat sich nun gezeigt, dass der ‚normale' Signifikanztest unerwünschte Eigenschaften hat. Diese theoretischen und praktischen Überlegungen zur Verwendung dieses Indizes bildeten die Motivation zur Erarbeitung des CS-Ansatzes.

Reflektiert kann und wurde der Signifikanztest hinsichtlich einiger Aspekte, welche wiederum auch verschiedenen Fachdisziplinen zugeordnet werden können: wissenschaftstheoretisch auf seinen logischen Status (z. B. Stegmüller, 1973), technisch-statistisch hinsichtlich der Fähigkeit adäquate Lösungen zu produzieren, soziologisch als Etablierung der positiven Wissenschaft (inkl. des ausgeübten Zwangs) und psychologisch hinsichtlich der Psychodynamik empirisch arbeitender Forscher. Hier sind diese Perspektiven jedoch nicht von Interesse. Fokussiert werden soll dagegen auf die Funktion in der Forschung.

Wird auf die Geschichte der empirischen Psychologie rekurriert, zeigt sich ein überraschendes Bild. Die wohl bekanntesten Namen der Psychologie wie Piaget, Köhler, Pawlov, Skinner und Bartlett haben keine Signifikanztests verwandt, haben sich sogar, wie etwa Skinner, offensiv gegen diese Praxis ausgesprochen. Die Einführung des Signifikanztests war in der Geschichte der Wissenschaften (siehe dazu besonders: Cowles, 1989; Danzinger, 1990; Gigerenzer, 2000) ein relativ spätes Ereignis, und auch die Psychologie konnte zum Zeitpunkt des Auftauchens des Signifikanztests auf eine kurze, aber reiche Vergangenheit empirischer Forschung zurückblicken. Diese Entwicklung begann Mitte des letzten Jahrhunderts, obwohl es weit früher schon Vorformen gab. So bewies etwa John Arbuthnot 1710 die Existenz Gottes durch einen Signifikanztest (auch wenn er dies nicht so nannte), später wurde dieser von Astronomen verwandt, um Beobachtungen zu identifizieren, denen sie nicht recht vertrauten (‚Ausreißer', siehe Gigerenzer, 2000). Wirkliche Verbreitung fand der Signifikanztest in der Psychologie erst durch das zweite Buch von Fisher (1935) (das erste Buch von Fisher, 1925, wurde wegen des hohen mathematischen Abstraktionsniveaus kaum in der Psychologie beachtet). Danzinger (1987, 1990) argumentiert hier, dass besonders ökonomische Zwänge, die Psychologie u.a. für die Pädagogik anwendbar zu machen, in Amerika diesen Trend förderten (in Deutschland bestand dieser Zwang vor dem 2. Weltkrieg

nicht).

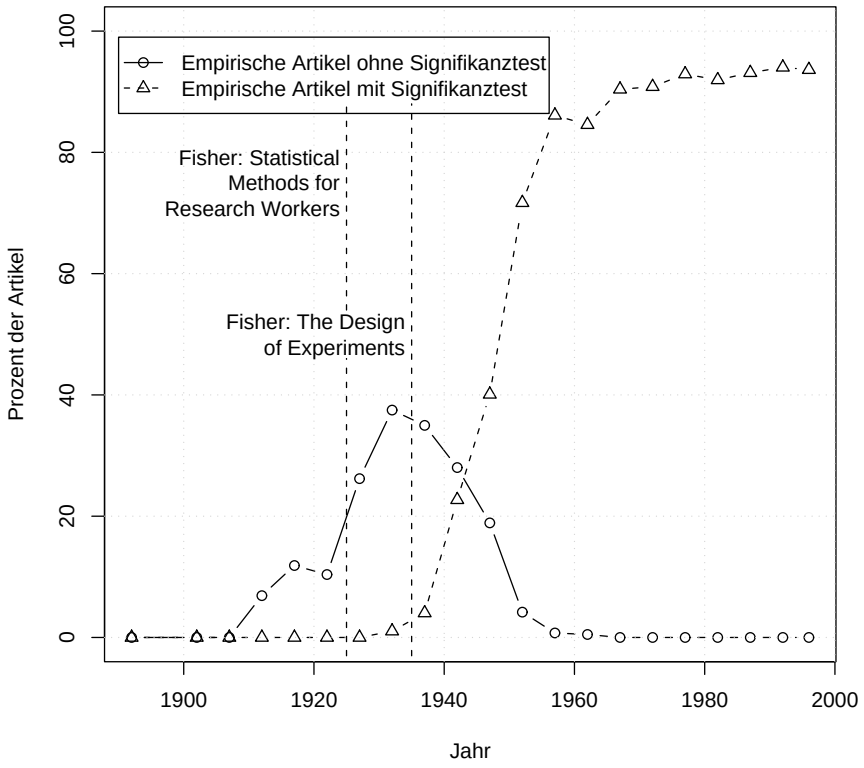

Abbildung 3.1: Historische Entwicklung des Signifikanztests in der Psychologie.

Abbildung 3.1 zeigt die prozentuale Anzahl der empirischen Arbeiten im Bereich Psychologie. Auf der Abszisse sind hierbei die Jahreszahlen, auf der Ordinate die relative Anzahl der Arbeiten wiedergegeben. Es wurden zweierlei Kurven eingezeichnet: Arbeiten mit (Dreiecke als Kurvenmakierung) und ohne Signifikanztest (Kreise als Kurvenmarkierung). Verwendet wurden, in leicht modifizierter Form, die in der Arbeit von Hubbard & Ryan (2000) wiedergegeben Daten. Als Orientierungspunkte wurden die Erscheinungsdaten der beiden Arbeiten von Fisher (1925, 1935) eingetragen. Diese Grafik zeigt deutlich den Anstieg des Signifikanztestes ab den 40er Jahren des 20ten Jahrhunderts. Die Grafik zeigt jedoch noch etwas anderes: Im Rahmen des generellen Anstiegs empirischer Arbeiten in der Psychologie *verdrängt* der Signifikanztest andere Verfahren. Diese bestehen zum großen Teil aus zusammenfassenden Maßzahlen, so wie es auch der ‚clinical significance' Ansatz darstellt. So kann der ‚clinical significance'-Ansatz auch als Beispiel der von Gigerenzer (2000) geforderten Flexibilität und Problemorientierung statistischer Verfahren genannt werden. So gesehen ist die Kritik am

Null-Hypothesen Test ein Regress in die Vergangenheit der empirischen Forschung, die ohne zwanghafte Schemata[1] ausgekommen ist.

3.1 Das Falsifikationsprinzip und der Signifikanztest

Nach der wohl von der Mehrzahl wissenschaftlich handelnder Psychologen als verbindlich angesehenen Wissenschaftstheorie Poppers ist eine Theorie bekanntlich entweder im Modus des ‚bis jetzt nicht widerlegten, also noch geltenden', oder als sich ‚als falsch erwiesenen'. Damit ist das positive Ganze der Wissenschaft eine Menge des hypothetisch Falschen; sicher dagegen ist einzig das sich als falsch Erwiesene. D. h. erweist sich nun eine Theorie als wirklich falsch (vorher war sie ja nur potentiell falsch), fällt sie aus der Menge des noch Geltenden heraus, wird Gegenstand historischer Erklärungen, aus denen sich etwa Fragen nach der Herkunft der jetzt geltenden Theorien ergeben, gerät in Vergessenheit oder wird als Abwehr der jetzt als irrig erkannten Meinungen angeführt. Diese letzte Funktion dieses negativen Wissensbestandes definiert also eine[2] Grenze des Korpuses des Noch-Bestehenden. Dieser Blick in das Forschungsgeschehen lässt den Signifikanztest als Instrument zu Verwerfung von Theorien *erscheinen* (dass er hier so nicht genügt, siehe unten).

Aufgrund der erwähnten Dimension des ‚noch-Geltens' kann neben dem Hinzukommen einer neuen Theorie — also der Vergrößerung des Korpus bestehenden Wissens — auch eine Zustandsänderung eines Mitgliedes dieser Menge ein Wissenszuwachs sein: Das neue Wissen kann also gerade darin neu sein, indem sich bisheriges Wissen als falsch erweist. Diese Falschheit kann bezüglich der sie betreffenden Theorie sowohl eine totale sein, so dass diese Theorie als Ganzes abgelehnt wird, als auch eine partielle, also bestimmte Teile der Theorie haben sich als falsch erwiesen und bedürfen einer Modifikation[3]. Interessanterweise (Žižek, 2001) widerspricht das Ergebnis der Linguistik (Reve & Busse, 1994), dass die meisten sprachlichen Regeln erst durch eine Ausnahme sich als Regeln überhaupt konstituieren dem Popperschen Falsifikationsprinzip: Die

[1] Die Forderung nach einer Flexibilisierung findet sich jedoch auch auf einer ganz anderen Ebene: Auf der Ebene der Schnittstelle zwischen Anwender und Verfahren. Dieser Shift, weg von der schematischen Auswertungen zur ‚greater statistics' (Chambers, 1999), findet sich auch auf Seiten statistischer Software wieder. Das Neuartige ist vor allem neben erleichternder funktionaler Programmiersprache, das Prinzip der Objektorientierung bei dem sowohl die Daten, als auch Ergebnisse und Prozeduren auf der abstraktesten Ebene das gleiche sind: Objekte. Ergebnis ist eine Flexibilität, welche der Datenanalyse einen neuen Charakter verleiht: den des Prozesses (Malewski & Oehlschlägel, 2000).

[2] Eine andere Grenze ist etwa diejenige zu den Theorien der so genannten Pseudowissenschaften. Aus diesem Bereich herrührende Anschauungen werden oft aufgrund grundsätzlicher Überlegungen abgelehnt. Hier wird also diese Grenze durch eine andere Prozedur erzeugt.

[3] Der zweite Fall wird gemeinhin auch als Exhausionsprinzip (Gadenne, 1984) bezeichnet, welches eine lange Diskussion hervorgerufen hat, ob die Poppersche Konzeption hier anwendbar sei. Der Widerspruch besteht darin, dass einmal eine Theorie als Vielheit, das andere Mal als Einheit verstanden wird. Das Poppersche Prinzip zielt auf *eine* Theorie, welche verworfen wird, während das Prinzip der Exhausion die Änderung eines Akzidenz der Theorie meint, also die Theorie als zusammengesetzt begriffen wird (Vielheit). Bemerkenswerterweise muss das Akzidens, welches spezifiziert wird, noch nicht einmal bei Formulierung der Theorie bekannt gewesen sein. Lösbar ist der Widerspruch, indem die Änderung eines Akzidenz einer Theorie eine neue Theorie genannt wird (Einheit der Vielheit).

Regel wird erst eine Regel *durch* eine Ausnahme. Beispielsweise ist dies bei der Größe der (europäischen) Geldstücke der Fall: Mehr Wert bedeutet gößere Münzen/Scheine (Regel), außer: das 5-Cent Stück, welches größer als das 10-Cent Stück ist (Ausnahme)[1]. Hier bewahrheitet sich das Sprichwort ‚keine Regel ohne Ausnahme'. Somit widerspricht die empirische Realität dem Popperschen Falsifikationsprinzip[2]. Während die erste Art des Wissenszuwachses, das Widerlegen des Bestehenden, nach Konvention des Signifikanztest die Methodik der Wahl *erscheint*, ist das Finden von Hypothesen und Theorien innerhalb dieser Wissenschaftsauffassung nicht definiert.

Der Signifikanztest ist zumeist eine Entscheidung zwischen zwei Alternativen[3]. Gewöhnlich werden die als Null- und Forschungshypothese beschrieben. So finden wir etwa bei Everitt (1998) die Definition:

H_0**:** „The ‚no difference' or ‚no association' to be tested (usually by means of a significance test) against an alternative hypothesis that postulates non-zero difference or association."

H_1**:** „The hypotheses against the null hypothesis is tested"

Jedoch erscheint die Zuordnung von Null- und Alternativhypothese des Signifikanztests nur schwer möglich. ‚Null' bedeutet ja auch ‚kein Effekt' oder auch ‚kein Zusammenhang'. Welcher Forscher hat wirklich eine Arbeitshypothese, die besagt, zwischen dem Konstrukt A und B gibt es *keinen* Zusammenhang? Dieser Fall ist eher selten. Ein Grund, warum Signifikanztest und Poppersche Logik nur selten kompatibel sind (Gigerenzer & Murray, 1987).

Anzumerken ist jedoch, dass die Zuordnung der statistischen Hypothese ‚kein Effekt' in den so genannten Äquivalenzstudien (Ebbutt & Frith, 1998; Wellek, 1994) nicht zutrifft. Beispiel kann etwa die Medikamentenzulassungsstudie sein, bei der es darum geht, zu zeigen, dass die Wirkung eines neuen Medikaments der eines schon zugelassenen Medikaments entspricht. Dies kann etwa dann sinnvoll sein, wenn das neue Medikament andere Nebenwirkungen hat und mithin für andere Anwendungsbereiche geeigneter erscheint. Jedenfalls finden wir hier genau die umgekehrte Hypothesenstruktur:

H_0 Die beiden Treatments unterscheiden sich.

H_1 Beide Treatments sind gleich (oder besser: *ähnlich*, d.h. es wird ein Bereich ‚der Gleichheit' definiert).

[1] Hier handelt es sich jedoch nicht um ein genuin linuistisches sondern ein der symbolischen Ordnung zugehöriges Beispiel.

[2] Nach Žižek (2001) ist diese Struktur notwendig, damit Regeln, um es in Hegelschen Begriffen auszurücken, vom bewusstlosen *an-sich* zum bewussten *für-sich* zu bringen. D.h. ohne Ausnahme ist eine Regel implizit, erst durch die Ausnahme stößt das Individuum auf sie. Dies ist analog zu Heideggers Differenz von Vorhandenheit und Zuhandenheit (Heidegger, 1923, Seite 70-71): Beim Hämmern ist der Hammer zuhanden, ganz in den Prozess gleichsam eingelassen. Erst durch die Störung, dass der Hammer nicht mehr funktioniert, wird er erst ein Ding ‚Hammer', wechselt also in den Modus der Vorhandenheit. D.h. auch hier ist ein negatives Element notwendig.

[3] Beispiele von Anwendungen mehrerer Alternativen werden in Lehmann (1997) beschrieben.

Dieser Exkurs zeigt, dass die Zuordnung von Forschungsinteresse zu der Hypothese, welche die Unterschiedlichkeit ausmacht, nicht selbstverständlich ist. Er zeigt aber auch an, dass eine reflektiertere Anwendung des Signifikanztests dazu beitragen könnte, den Mangel der ‚Null' -Hypothese zu bereinigen. Hier sei jedoch nicht nur auf die Äquivalenztests hingewiesen, sondern vor allem auch auf die Möglichkeit, verschiedene Modelle gegeneinander zu testen. D.h., der Mangel des Signifikanztests ist wenigstens in dieser Hinsicht auch ein Mangel der Praxis.

3.2 Theoretischer Status des Signifikanztests

Dem Erfolg im Forschungsbetrieb steht paradoxerweise dessen theoretische Nichtexistenz der zumeist *angewandten Form* des Signifikanztests (siehe Abschnitt 3.3 auf Seite 31) gegenüber. Dies bedeutet nicht, dass dieser nicht theoretisch fundiert ist, sondern dass der gewöhnlichen Verwendung des Tests und mithin die Rolle innerhalb der Forschungslogik ein theoretischer Mangel anhaftet. Dieser gründet in der Differenz zwischen der theoretischen Konzeption (mathematischen Theorie) und praktischer Anwendung. Vielmehr ist der zur Anwendung gebrachte Signifikanztest eine Mischung aus unterschiedlichsten Test-Theorien stammender Ideen. Gigerenzer bezeichnet diese Mischung verschiedener Theorien als ‚hybrid' (Gigerenzer, 2000; Gigerenzer & Murray, 1987; Gigerenzer, 1989a, 1998): und beschreibt den herrschenden Zustand anhand des zweiten Strukturmodells von Freud(Gigerenzer, 2000):

Über-Ich: verlangt, dass nach Neymans und Pearsons Theorie gearbeitet wird. Nach diesen Theorien sind zwar längst nicht alle der verwandten Tests konstruiert, jedoch herrscht die Vorstellung vor, dass es ‚eigentlich' richtig ist, eine Entscheidung mit Abwägung des Fehlers der ersten *und* zweiten Art zu treffen.

Ich: erinnert daran, was alle tun: Nur einen Fehler zu beachten. Dies ist insbesondere bei Fisher's Null Hypothesen Tests der Fall, bei denen ausschließlich der Fehler erster Art eine Rolle spielt. In der Praxis wird jedoch über die Null-Hypothese nicht nachgedacht und es wird alleine auf die Alternativhypothese fokussiert.

Es: Wunschergebnis eines statistischen Tests ist die Wahrscheinlichkeit der Hypothese unter der Bedingung der erhobenen Daten $P(Modell|Daten)$ statt $P(Daten|Modell)$. Dieser Sachverhalt ist auch der geläufigste Fehler in der Interpretation des Signifikanztests (siehe unten). Im Grunde handelt es sich um einen Wunsch nach einer Bayes-Statistik.

Es ist eine historische Fußnote, welche gleichwohl in das Lagerdenken dieser Zeit[1] passt, dass die Begründer der Theorien, besonders Fisher und Neyman, sich erbittert, auch auf einer persönlichen Ebene gegenüber standen. Doch auch im Kern des zugrunde liegenden Gedankenganges unterscheiden sich die Konzeptionen radikal voneinander. Neyman und Pearson bezeichnen ihre Konzeption selbst als ‚behavioural', d. h. auf

[1] vgl. etwa auch der Streit von Behaviorismus und Psychoanalyse bis in unsere Zeit, welcher sich erst mit neueren Arbeiten Grawes anschickt, sich in einem höheren allgemeineren Standpunkte aufzulösen.

konkretes Handeln bezogen. So ist ihr paradigmatisches Beispiel das einer Fabrik, die eine Ware herstellt. Ausgangspunkt ist, dass die herausgehende Ware in Ordnung ist (Nullhypothese). Es werden jetzt kleinere Stichproben gezogen, um zu entscheiden, ob der Produktionsprozess angehalten werden muss, da zu viele schadhafte Waren produziert werden. Dabei ist der Fehler erster Art, den Produktionsprozess fälschlicherweise zu unterbrechen. Jedoch weiß jeder Unternehmer, wie gefährlich es ist, schlechte Waren zu produzieren und damit den Verbraucher zu verärgern. Dies führt auf den Fehler zweiter Art: den Produktionsprozess fälschlicherweise nicht zu unterbrechen und einen Fehler zu übersehen. Eine Abwägung, welcher Fehler bedeutsamer ist, erfolgt hierbei durch die hypothetischen Kosten, die mit diesen Fehlern verknüpft sind.

Paradigmatische Abbildung dieses Verhältnisses zwischen zweierlei Arten von Fehlern sind zwei sich überlappende Verteilungen, so wie dies auch in Abb. 2.1 auf Seite 9 geschehen ist. Interessanterweise werden derartige Überlegungen zum Fehler erster und zweiter Art kaum angestellt und es wird blind auf die Konvention $\alpha = 0.05$ (der Fehler zweiter Art wird ‚verdrängt') gesetzt. Fisher hat diese Konzeption immer als ‚nicht-wissenschaftlich' verworfen. Seine Konzeption umfasst allein den Fehler erster Art. Abb. 2.2 auf Seite 13 kann hierbei als Illustration dieser Konzeption verstanden werden. Interessanterweise hielt Fisher selbst den Signifikanztest für ein ‚weak argument' für die Forschung.

In der Analogie zur neurotischen Verarbeitung von Gigerenzer (2000) persistiert der Konflikt der Eltern in der gegenwärtigen Situation, wird jedoch nicht mehr erinnert, sondern führt zu einem permanent schlechten Gewissen des Forschers und zu ‚rituellem Händewaschen'. Die Praxis sieht folgendermaßen aus: Es wird nur eine Alternative in der Formulierung der Hypothese beachtet und ein Abwägen des Fehlers erster und zweiter Art bleibt aus. Mit der Formulierung *einer* Hypothese befindet sich die ‚hybride Logik' in der Nähe der Fisherschen Theorie, während Neyman-Pearsonsche Prüfverfahren angewendet werden. Indem weiterhin die p-Werte als Wahrscheinlichkeit der Hypothese interpretiert werden, in der bayesianischen Logik. ‚The results are wishful thinking, suppression of conflicts, and a statistical practice – null hypothesis testing – that resembles ritualistic handwashing'

Wie gezeigt, verbergen sich in der geschichtlichen Entwicklung nicht thematisierte Entscheidungen wie etwa:

1. Ob die Daten im Licht der Theorie zu betrachten seien oder die Theorie im Licht der Daten.

2. Ob in der Forschung eine diskrete Entscheidung oder ein kontinuierlicher Wissenszuwachs angestrebt wird.

Die erste genannte Unterscheidung ist im Kontext des etablierten Hypothesentestens als gängiger Fehlschluss bekannt: P-Werte werden als Wahrscheinlichkeit der Hypothese unter der Bedingung der Daten interpretiert, also etwa ‚Im Lichte unserer Daten konnte unsere Theorie mit 95% Wahrscheinlichkeit verifiziert werden'. So einleuchtend diese Interpretation für einen naiven Zuhörer ist, so falsch ist sie auf dem Hintergrund

der statistischen Theorie, mit der gewöhnlich die Daten ausgewertet werden. Diese drückt nämlich genau das Gegenteil aus, d.h. statt der einleuchtenden

$$P(Hypothese|Daten) \tag{3.1}$$

Wahrscheinlichkeit der Hypothese unter der Bedingung der Daten liefert diese die

$$P(Daten|Hypothese) \tag{3.2}$$

Wahrscheinlichkeit der Daten unter der Bedingung der Hypothese aus. Dem verwunderten ‚naiven Betrachter' wird entgegnet, dass sich Daten an sich nicht theorielos sammeln lassen, also immer unter der Bedingung der Theorie zu betrachten seien. Ein historisches Beispiel hierfür, ist etwa die Situation der Astronomen vor einhundert Jahren. Diese wollten herausfinden, welche Beobachtungen sie als unwahrscheinlich verwerfen wollten: Ein klassisches und sehr sinnvolles Anwendungsbeispiel des Signifikanztests. Jedoch ist der Forscher meist an etwas ganz anderem interessiert: Wie steht es um seine Theorie?

Die Testtheorie, welche diesen Mangel beheben würde, wird allgemein als subjektivistisch abgelehnt: Die Bayes-Theorie[1]. Hier muss vor dem statistischen Test der so genannte ‚prior believe', das (nicht ganz wörtlich) als ‚vorhergehendes Wissen' übersetzt werden kann, angegeben werden[2]. Durch die Daten wird dieses vorherige Wissen nur korrigiert. Ist recht wenig vorheriges Wissen vorhanden, wird das Ausmaß der Korrektur recht groß ausfallen, ist recht viel bekannt, recht klein. Es handelt sich also im Gegensatz zu der klassischen Statistik weniger um ein ja/nein, als um die Schätzung der Wahrscheinlichkeit, auch für ein ja/nein. Häufigstes Argument gegen diese Art der Statistik ist, dass die Wahl der Grundwahrscheinlichkeit ein subjektives Element beinhaltet. Dies liegt an der Tatsache, dass das vorherige Wissen mit in die Analyse eingeht, diese mithin nicht unabhängig vom Eingriff des Anwenders ist und somit die statistische Analyse ihre Objektivität verlöre. Die Entgegnung, dass jede statistische Analyse nie ganz objektiv sein kann, sondern immer auch ein subjektives Element beinhaltet, trifft hier nicht ganz, da zugegebenerweise der Spielraum bei einer nach der Bayes Statistik vorgehenden Analyse größer ist. Allerdings ist das subjektive Element kleiner als die Gegner der Bayes-Statistik annehmen, da eigentlich weniger der ‚Glauben' denn das vorhergehende Wissen in die Analyse eingeht. Ist dieses nicht vorhanden, kann dies mit einer informationslosen ‚prior' ausgeglichen werden, der Gleichverteilung. Konsequenz aus der auf der Bayes Statistik beruhenden Forschung wäre auf jeden Fall,

[1] Zur Bayes-Statistik sind sehr viele gute Lehrbücher verfügbar. So geben Berry (1996), Abrams, Ashby, & Errington (1994) und Spiegelhalter, Freedman, & Parmar (1993) einen grundlegenden Einblick in die Bayes-Statistik, während Gelman, Clarin, Stern, & Rubin (1997) eher auf einem höheren Niveau in diese einführen. Die sehr lesenswerten Beiträge von Edwin T. Jaynes zur Bayes Statistik, welche sich sowohl mit der Abgrenzung zur klassischen Statistik (Jaynes, 1984, 1990, 1985) beschäftigen und auch zur Fundierung der Forschung als ganzes beitragen (Jaynes, 2003) sind auch im Internet verfügbar: `http://bayes.wustl.edu/`.

[2] Bei ‚emirical bayes' Verfahren wird der prior durch die Daten selbst geschätzt. Dies ist etwa bei den hierarchischen Analysen der Fall. Gelman et al. (1997) kritisieren diesen Begriff mit Recht, da eigentlich jedes Bayes-Verfahren empirisch ist.

dass bisherige Ergebnisse als ‚Prior' genommen werden, so dass sich sukzessiv von Studie zu Studie eine bessere Schätzung des Effektes ergibt. Dies hätte unter anderem den Vorteil, dass die letzte Studie zu einer bestimmten Forschungsfrage automatisch den geltenden Stand der Forschung wiedergeben würde. Diese Konzeption vermeidet auch die immensen Probleme, die sich bei der Integration von Ergebnissen ergeben, die alleine Signifikanztests beinhalten.

3.3 Praktische Konsequenzen des Signifikanztests

Abbildung 3.2 illustriert das Hauptargument bzgl. der praktischen Schwierigkeiten bei der Anwendung von Signifikanztests: Je größer die Stichprobe, desto eher wird auch ein kleiner Effekt ‚signifikant'. Auf der Abszisse der Abbildung 3.2 ist die Power des

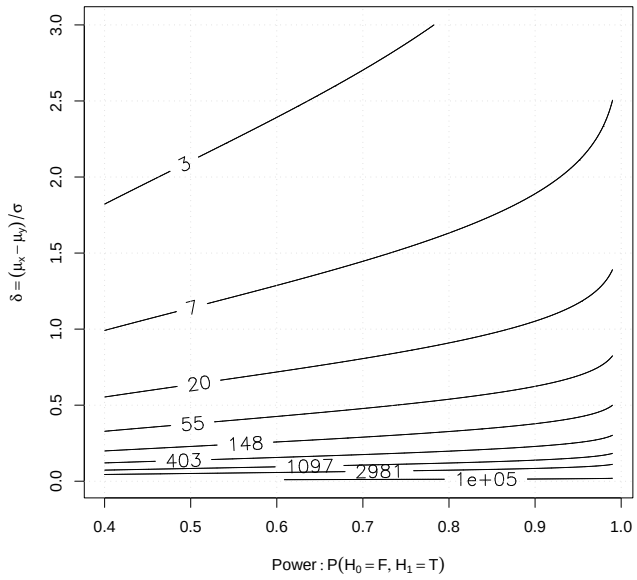

Abbildung 3.2: Zusammenhang zwischen der Power und Stichprobengröße bei einem T-Test.

Tests, also die Wahrscheinlichkeit, den Unterschied zu entdecken, abgetragen. Die Ordinate gibt die normierte Mittelwertsdifferenz (oft fälschlicherweise als Effektstärke bezeichnet) beider Gruppen des T-Tests wieder. Die Linien repräsentieren bestimmte Stichprobengrößen. So bedeutet, dass bei einer Stichprobengröße von $N = 3$ mit einer Power von 0.7 eine normierte Mittelwertdifferenz von ca. 2.7, dagegen mit einer Stichprobengröße von $N = 2981$ eine normierte Mittelwertdifferenz, die sehr nahe 0 ist, entdeckt werden kann.

Dies führt unweigerlich zu der Frage, was dieser kleine Effekt denn eigentlich bedeutet — ein Hauptargument, eine inhaltliche Bedeutung zu fordern, wie dies etwa in den ‚clinical significance' Ansatz beispielhaft geschieht. In der wirklichen Forschung ist jedoch oft ein gegenteiliger Effekt zu finden: Mit Vergrößerung der Stichprobe ‚verschwinden' die Effekte. Dies steht nicht im Widerspruch zu der obigen Aussage, sondern beruht auf einem ganz anderen Effekt. So sind besonders auf Verteilungsannahmen beruhende statistische Verfahren oft sehr empfindlich gegenüber ‚Ausreißern', die oft Pseudoergebnisse produzieren. Auch ist es leider oft die Praxis, dass sehr viele Tests gerechnet werden, bis einer von diesen ‚signifikant' wird. Ein dritter Mechanismus, der bei kleineren Stichproben oft unterschätzt wird, ist der des ‚Overfit'.

Um noch einmal die wichtigsten Argumente, die gegen die Verwendung des Signifikanztests sprechen und teilweise nicht genannt wurden, aufzuführen, wurde die folgende Übersicht (siehe 3.3 auf der nächsten Seite) erstellt.

3.4 Rückbindung des Exkurses

Die statistische Methodik ist somit in ihrem geschichtlichen Gewordensein mit bewusst oder unbewusst gefällten Entscheidungen verbunden. Dies bedeutet jedoch nicht, dass das statistische Handeln willkürlich ist, sondern dass dieses immer wieder hinsichtlich der forschungslogischen Zielsetzung reflektiert werden muss (Kordy, 1986). Empfundene Mängel an den Ergebnissen des statistischen Schließens, insbesondere die Frage nach der Bedeutsamkeit der Ergebnisse dieses Verfahrens, war Hauptmotivation zur Entwicklung des CS-Konzepts. Die Reflektion dieser Entwicklung zeigt aber, dass auch dieser Index nicht ein factum brutum ist, sondern selbst immer wieder bzgl. seiner Anwendung reflektiert werden muss. Die vorliegende Arbeit kann als Beispiel zu diesem veränderbaren Charakter der Methodik gelesen werden.

- Die Nullhypothese ist gewöhnlich an sich nicht von Interesse, sondern wird eher aus mathematischer Bequemlichkeit gewählt.
- Auch ein sehr kleiner Effekt kann bei einer hinreichend großen Stichprobe nachgewiesen werden.
- Der Cut-Off von 0.05 als Grenze von ‚signifikant' vs. ‚nicht signifikant'ist ein Resultat des persönlichen Kleinkrieges zwischen Fisher und Pearson (Cowles, 1989). Pearson verweigerte Fisher andere Tabellen als für $\alpha = 0.05$. Später schreibt Fisher: ‚No scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; rather he gives his mind to each particular case in the light of his evidence and his ideas' (Fisher, 1956). Und empfiehlt (wie übrigens auch Neyman), das exakte Signifikanzniveau zu berichten (z.B. $p = .03$, nicht $p < .05$)(Gigerenzer, 1998). Allerdings besteht bzgl. dieses Punktes längst keine Einigkeit, so finden sich bei Fisher Stellen, in denen er genau das Gegenteil fordert.
- P-Werte werden oft fehlinterpretiert:
  - P-Werte werden als Wahrscheinlichkeit der Statistik interpretiert.
  - P-Werte werden irrtümlich als die Irrtumswahrscheinlichkeit, dass die Forschungshypothese zutrifft, verstanden.
- Fragen, wie die der Identität zweier Behandlungsverfahren sind schwer durchführbar.
- Bei der Integration von Ergebnissen (multiple comparisions, Metaanalysen) hat die klassische Statistik sehr große Schwierigkeiten. Hier sind reine Bayes-Ansätze eindeutig im Vorteil.
- P-Werte können nur gegen eine Hypothese sprechen, niemals für ihre Gültigkeit.
- Die Testprozedur ‚vergisst' das vorher bekannte Wissen. Dies ist auch ein Einwand, der aus dem bayesianischen Denken entspringt: Hier könnten vorhergehende Ergebnisse in die aktuelle Untersuchung integriert werden, d.h. die letzte Untersuchung zu einem Thema spiegelt immer den aktuellen Forschungsstand wieder.
- Die Bestimmung der Stichprobengröße bei der Studienplanung ist oft willkürlich und wird oft mehr durch wirtschaftliche Zwänge bestimmt.
- Designs mit einem festen N sind oft unbefriedigend. So ist es häufig so, dass bei dem Ende einer Studie noch keine wirkliche Aussage hinsichtlich des Effekts möglich ist, so dass es viel sinnvoller wäre ‚weiter zu machen'(was macht man mit einem Ergebnis von $p = 0.06$?). Umgekehrt könnten Studien auch verkürzt werden, wenn genügend Daten da sind, um eine Aussage treffen zu können.
- Was bedeutet ein signifikanter P-Wert bei einer sehr großen Stichprobe, und: was bedeutet ein nicht-signifikanter P-Wert bei einer kleinen Stichprobe?
- P-Werte sind Wahrscheinlichkeiten, bei Wiederholung das vorliegende oder ein noch extremeres Ergebnis zu erhalten. Das heißt, dass die Ergebnisse interpretiert werden als ‚what would have occurred following results that were not observed at analyses that were never performed' (Emerson, 1995)
- Wie ist damit umzugehen, wenn sehr viele Tests durchgeführt werden (alpha-Adjustierung)? Die verbreitete Praxis der alpha-Adjustierung beruht auf fragwürdigen Voraussetzungen, so etwa, die das a priori alle Hypothesen die gleiche Wahrscheinlichkeit ($\frac{1}{2}$) haben. So schreibt Rothman (1990): „The theoretical basis for advocating a routine adjustment for multiple comparisons is the 'universal null hypothesis' that 'chance' serves as the first-order explanation for observed phenomena. This hypothesis undermines the basic premises of empirical research, which holds that nature follows regular laws that may be studied through observations. A policy of not making adjustments for multiple comparisons is preferable because it will lead to fewer errors of interpretation when the data under evaluation are not random numbers but actual observations on nature. Furthermore, scientists should not be so reluctant to explore leads that may turn out to be wrong that they penalize themselves by missing possibly important findings." (vgl. auch Cook & Farewell, 1996)

Tabelle 3.1: Übersicht zu einigen Argumenten bzgl. des Signifikanztests

4 Exkurs: Normalität als Diskursdispositiv. Jürgen Links Versuch über den Normalismus

Fernes zu erkennen ist meist einfach, das Nächstliegende, oder Nahe, ist meist schwer benenn- und erkennbar. Das uns Nahe ist, überraschenderweise, der Begriff der Normalität. Link (1999) hat die Wirkung dieser Begrifflichkeit durch viele Beispiele des Alltagslebens aufgezeigt. Dass diese Begrifflichkeit auch für die vorliegende Arbeit von immenser Bedeutung ist, bedarf die Darstellung keiner weiteren Rechtfertigung, jedoch ein gewisses Maß an reflexiver Rückbindung.

Kant hatte nur drei Fragen, die er sich stellte: Was kann ich wissen? Was darf ich hoffen? und Was soll ich tun? Diese entsprachen den drei Kritiken (Kritik der reinen Vernunft, der Urteilskraft und der praktischen Vernunft), die er schrieb. Bekanntlich ist eine seiner Antworten (Link, 1999) der kategorische Imperativ: ‚handle nur nach derjenigen Maxime, durch die du zugleich wollen kannst, dass sie allgemeines Gesetz werde'. Welches ist nun der Unterschied zur ‚Normalität'? Dieser besteht darin, dass es sich bei dem Kantischen Satz um eine ‚Norm' handelt, ‚normalistisch' dagegen könnte man diesen etwa so umformulieren: ‚Handle so, wie es normal ist'. Bevor jedoch weiter auf diese Unterscheidung eingegangen wird, an dieser Stelle erst einmal einige Anmerkungen zur eigentümlichen, ‚disursanalytischen' Methode.

4.1 Funktion im Diskurs

Als Diskurse können Systeme von Aussagen bezeichnet werden, die das gesellschaftliche Handeln strukturieren und organisieren. Diskurs*tragende* Kategorien sind weiter solche, die würde man sie aus dem Diskurs entfernen, diesen in sich zusammenstürzen ließen (Link, 1999, Seite 15). Es handelt sich also um notwendige Bedingungen des Diskurses. Es zeigt sich nun, dass Diskurse, gerade auch weil sie sich in verschiedene ‚Spezialdiskurse' vereinzeln (z.B. politische, wissenschaftliche, juristische, medizinische Diskurse), neben den tragenden Kategorien auch Anknüpfungspunkte, die die verschiedene Diskurse miteinander verbinden, benötigen. Eine solche diskurstragende und als Interdiskurs verbindende Funktion ist die Begrifflichkeit der Normalität. Jeder Interdiskurs, mithin auch der der Normalität, stellt einen bestimmten Bildraum bereit, auf den der Spezialdiskurs per Analogie bezogen werden kann, und er bezieht sich dabei nicht nur auf einen Spezialdiskurs, sondern sogleich auf mehrere. Er steht als ein

verbindendes Glied zwischen den Diskursen. Dabei geht es nicht nur um eine irgendwie zu leistende Integration der Diskurse, sondern zugleich um die Durchsetzung eines Sinnschemas, das bestimmte Logiken in die Spezialdiskurse hineinprojiziert.

4.2 Geschichtliche Entwicklung

Von ‚Normalität' kann erst seit dem 18. Jahrhundert sinnvoll gesprochen werden. Um 1800 konstituieren sich die später als Humanwissenschaften bezeichneten Fächer und es vollzieht sich der Wechsel vom ‚taxonomischen' zum ‚historizistischen' Denken. Michel Foucault hat das eingehend dargestellt (Foucault, 1990). Zuvor jedoch herrschen protonormalistische Begriffspaare wie ‚Gleichgewicht/Ungleichgewicht', ‚Identität / Nicht-Identität', ‚gesund/krank', ‚legitim / illegitim', ‚schön / hässlich', ‚natürlich / unnatürlich' usw. vor. Nicht *Normal*ität ist hier der Bezugsrahmen, sondern *Normativ*ität.

Im 19. und 20. Jahrhundert tritt diese Begrifflichkeit in den Hintergrund und die Epoche des ‚Normalismus' beginnt. In dieser Epoche breiten sich aus der institutionalisierten Statistik und Empirie bestimmte Vorstellungen und Bildlichkeiten (so die Glockenkurve) aus. Die damalige Psychologie hat jedoch nach Link wenig dazu beigetragen, war sie doch in weiten Fällen ‚typologisch'. Sie hat, ebenso wie der Spezialdiskurs der Psychoanalyse, nur einen geringen Anteil an der Herausbildung des Normalismus. Die Diskontinuität, Heterogenität und Pluralität der vielen ‚Typen' von ‚Geisteskrankheiten' widerspricht den Prämissen des Normalismus. Die Geschichte der Normalität bekommt erst bei Sigmund Freud eine neue Wendung. ‚Wir glauben nicht mehr', schreibt er in seinem Aufsatz ‚Eine Kindheitserinnerung des Leonardo da Vinci' (Freud, 2000), ‚dass Gesundheit und Krankheit, Normale und Nervöse, scharf voneinander zu sondern sind.' Gleichwohl lässt diese protonormalistische Begriffsverwendung darauf schließen, dass Freuds Psychoanalyse im strikten wissenschaftlichen Sinne in die Entfaltung des Normalismus nicht integrierbar ist. Bei Freud, argumentiert Link, gerate das Normale ‚in Anführungszeichen'. Beispielhaft wird der Beitrag der Psychoanalyse zum flexiblen Normalismus anhand der amerikanischen ‚Therapiekultur' beschrieben, die eine ‚Therapierung der Normalen' betreibt und - im Unterschied zum Protonormalismus mit seiner eher auf ‚Außensteuerung ' gerichteten Begrifflichkeit tendenziell die gesamte Gesellschaft erfasst. Link referiert hier auf Robert Castel, der die Aufgabe der Psychoanalyse darin sieht, ‚die beiden grundsätzlichen Anforderungen an eine moderne Politik der geistigen Gesundheit' zu übernehmen: erstens den absoluten Charakter des Einschnitts normal/pathologisch der alten psychiatrischen Symptomatologie aufzubrechen, um in den Übergangszonen greifen zu können, wo die Grenzen von Anormalität und mangelnder sozialer Anpassung sich verwischen und zweitens die Langzeitproblematik in den Griff zu bekommen, das Risiko einer Pathologie während der Latenz zu antizipieren, um den Übergang von der Repression a posteriori zur Prävention a priori zu gewährleisten.

Norm und Normalität

Normativität	**Normalität**
(juridiforme, binäre) Erfüllungsnorm	Orientierungskarte (für Manövrieren/ Adjustieren) (=Orientierungsnorm)
Abweichung 1 (Normbruch)	Abweichung 2 (Exploration)
(binäre) imperative Punktnorm	um den Durchschnitt situierter ‚normal range' mit Normalitätsgrenzen (=Grenz-‚norm', Schwellen-'norm')
präskriptiv	deskriptiv
Norm (bezüglich individuellen Handelns)	Tendenz (kollektiven Handelns)
juridiforme Dispositive mit Sanktionen	statistische Dispositive von Risiko / Versicherung
dem Handeln (idealtypisch) präexistent	dem Handeln idealtypisch postexistent
(unabhängig von Verdatung)	notwendig gestützt auf Verdatung als "historisches Apriori"
panchronisch-transhistorisch	ausschließlich modern

Tabelle 4.2: Überblick zu den Begriffen Norm und Normalität (Link, 1999, S.444)

4.3 Normativität vs. Normalität

Wie schon erwähnt, muss man grundsätzlich zwei Kategorien unterscheiden, die im (Sprach-)Gebrauch oft vermischt verwendet werden, jedoch theoretisch-semantisch strikt zu trennen sind: Normativität vs. Normalität. Diese beiden zu unterscheidenden Kategorien lassen sich gut am Beispiel der Diskussion des Schwangerschaftsabbruchs erläutern: Während eine religiöse oder klassisch-juristische ‚Norm' ein bestimmtes Verhalten (z.B. Abtreibung) eindeutig als entweder zulässig oder unzulässig definiert, stellt sich die (‚alltägliche') Frage nach der ‚Normalität' von Abtreibung völlig anders. Sie lautet, ob etwa das tatsächliche Verhalten als ‚tolerabel' und ‚akzeptabel' gelten kann, oder ob dringend interveniert werden muss.

Die aus Link (1999, Seite 444) entnommene Tabelle 4.2 gibt einen Überblick zu den beiden Konzepten. Die Kategorie der *Norm* umfasst das semantische Feld ‚normativ', ‚Rechtsnorm', ‚Normativität' etc. Hierbei handelt es sich um (quasi-juristische) Normen, die eindeutig festgelegt sind. Durch sie lassen sich ‚richtig' und ‚falsch' klar voneinander trennen, denn Normen legen binäre Gegensatzpaare fest. Da sie Konstitutens menschlicher Gesellschaften sind, sind sie somit immer existent. Verstöße gegenüber diesen Normen sind Normbrüche, d.h. potentiell ist ein Gesetz gebrochen worden. Dem gegenüber gibt es das Feld *normal*, ‚Normalität', ‚Normalisierung' etc. Hier sind nicht zwei unverbundene, binäre Kategorien wie ‚richtig' oder ‚falsch' gegeben, sondern es liegt eine kontinuierliche Größe vor. Wichtig für die Kategorie ‚normal' ist,

dass sie nicht auf sicheren Fakten, sondern auf Schätzungen bzw. auf der Antizipation eines potentiell faktensetzenden Konsenses basiert. Link bietet folgende Definition an: ‚Betrachtet als normal, was von anderen als normal betrachtet werden könnte' (Link, 1999, Seite 16). In dieser Definition wird die prognostische Komponente, die Hochrechnung deutlich, die gleichzeitig eine Unsicherheit bedeutet. Darüber hinaus ist auch der Verweis auf die anderen als notwendige Komponente von Normalität enthalten. Hier deutet sich auch schon die in den Individuen vorhandene Tendenz der Selbstnormalisierung, via Identifikation mit einem virtuellen anderen[1], an.

Normalisierung ist nun der Versuch, ein Phänomen in ein Feld vergleichbarer Phänomene einzureihen. Voraussetzung dafür sind Homogenisierung - denn erst diese stellt die Vergleichbarkeit her - und Niedrigdimensionalität der Phänomene. Es wird ein Kontinuum zwischen dem ‚Normalen' und dem ‚Anormalen' angenommen, ‚Normalität' ist folglich stark dynamisch. Somit ist die Frage, was ‚normal' ist, veränderlich und auch abhängig vom sozialen Kontext. Die Grenzen der ‚Normalität' werden immer wieder neu gezogen. Dieser Prozess ist jedoch sehr problematisch, da in seiner Festlegung ein Zirkelschluss wirksam ist: ‚Normal ist, was normalerweise als normal gilt.' Zusammenfassend lässt sich sagen, dass Normalität das Resultat von Normalisierung ist und vom System produziert wird - dem System des Normalismus.

Welches ist und war die Quelle dieser Begrifflichkeit? Dabei kann der Normalismus als ein Reflex auf moderne Dynamiken, die oft exponentielle Trends zeigen, verstanden werden. Der Normalismus dient der Regulierung und der Kompensation dieser Trends, sowie der Stabilisierung des kollektiven und des individuellen Systems (Link, 1999, Seite 313). Nach Link haben gerade die exponentiellen Dynamiken des modernen Fortschritts den Normalismus hervorgebracht, der somit ein Phänomen des Industriezeitalters ist. Damit wird auch die historische Spezifität des Normalismus deutlich. Es handelt sich um ein sozio-kulturelles Phänomen und nicht um ein biologisches oder anthropologisches. Dass dabei biologische u.a. Disziplinen auf Schritt und Tritt auf Normalitäten stoßen, kann wiederum als Wirkung dieser Begrifflichkeit verstanden werden. Historisch gesehen haben sich Normalismus und Normativität seit dem 18. Jahrhundert zunehmend unabhängig voneinander entwickelt. In den Mechanismen des Herrn Jedermann spiegelt sich diese Begrifflichkeit, getrieben als Angst und Unsicherheit wieder, es kann von einer ‚Denormalisierungsangst', d.h. der Angst irreversibel denormalisiert zu sein, gesprochen werden. Dies ist die größte Angstquelle im Normalismus (Link, 1999, Seite 61 und S.263).

Der Normalismus soll gerade versichernd wirken, indem die exponentiellen Wachstumstrends reguliert werden. Dabei werden die exponentiellen Kurven quasi ‚zurückgebogen' auf Gaußkurven. Diese Gaußkurven entsprechen den verinnerlichten Normalfeldern, die eben gerade normalisierend wirken. Die Statistik wird somit als eine willentlich eingesetzte, intervenierende Taktik (Verdatung) verwendet und nicht als deskriptive Wissenschaft. Die kollektiven und symbolischen Normalfelder wechselwirken mit bereits bestehenden Ebenen, z.B. mit juristischen Normen. Der Normalismus etabliert parallel dazu eine zweite Ebene, die das Verhalten rein statistisch erfasst

[1] Bei Lacan als der große Andere bezeichnet (Žižek, 2001).

und, zwischen zwei Polen verteilt, anordnet. Wenn es zu starken Abweichungen zwischen dem festgeschriebenen, expliziten und dem ‚spontanen' Willen der Gesellschaft kommt, entsteht ein großes Risiko der Denormalisierung. Dieses muss ausgeglichen werden, was durch eine Adjustierung der Normen geschehen kann, die 'alte' Ebene wird normalisiert (Seite 344).

Das normalistische Leben ist somit ein Leben in einer Kurvenlandschaft aus ‚Signal- und Kontrollebenen' mit dem Ziel, die Denormalisierungsangst endgültig zu bewältigen. Dies ist jedoch utopisch, da ‚im Schatten der versichernden Denormalisierungsangst um so mehr jene heimliche Denomalisierungssehnsucht anwächst, die von der entropischen Tendenz der Intensitäten gespeist wird und deren Ambivalenz ihrerseits die Denormalisierungsangst speist' (Seite 345).

Nach Link können zwei Strategien zur Erzeugung von Normalität bzw. zwei Antworten auf die Denormalisierungsangst unterschieden werden (Link, 1999, Seite 75ff):

Fixistischer Protonormalismus Die Normalität wird mit allen Mitteln verteidigt. Ziel ist es - z.B. im Falle von Spannungen zwischen der Ebene der Normen und der des statistischen Verhaltens - den status quo ante herzustellen. Die Normalitätsgrenzen werden fixiert, um Normalität zu entdynamisieren und auch zu enthistorisieren.

Flexibler Normalismus Normalitätsgrenzen sind flexibel verschiebbar. Dies bewirkt, dass das System veränderbar ist. Die Funktion ist klar: flexible Systeme können sich anpassen und sind hinsichtlich innerer und äußerer Konflikte stabiler. Entwickelt hat sich diese Strategie gegen Ende des 19. Jahrhunderts, als der Protonormalismus die auftretenden starken Dynamiken nicht mehr kompensieren konnte. Historischer Einschnitt ist hier insbesondere die so genannte ‚68-Bewegung'.

4.4 Kollektivsymbolik

Mit der Normalität verbunden ist ein ganzes Ensemble von Bildlichkeit. Link nennt diese Summe aller, mit dem Begriff der Normalität verbundener Bilder ‚Kollektivsymbolik'. Diese ist hinsichtlich der ‚Normalität' konstituierend. Beispiele solcher Symboliken sind Ausdrücke wie ‚ein Boot, in dem wir alle sitzen', seine diversen ‚Wenden', ‚Schieflagen', ‚Auf- und Abschwünge', die Kopplung des ‚Bootes' mit den Bildern ‚Auto', ‚Konjunkturmotor stottert' , ‚Flugzeug' (‚Turbulenzen bei der Finanzierung des Gesundheitswesens') und auch ‚Körper' (etwa ein Mensch, dessen Belastbarkeit für Abgaben und Steuern, wie das medizinische Testen der Schmerzgrenze, funktioniert).

Dieses Arsenal an Bildlichkeit ist im alltäglichen Diskurs an normalisitische Kurven repräsentiert, wobei die Allgegenwart dieser Begrifflichkeit sicherlich auch zu deren Unscheinbarkeit beiträgt. Verdeutlicht man sich, dass statistische Kurven und Verteilungen kontinuierlich sind und keine fixen Grenz-Markierungen aufweisen, dann müssen die Grenzwert-Einschnitte (‚nicht mehr als 100.000 Flüchtlinge im Jahr' ‚kein Dollarkurs unter 1,28 DM', ‚nicht mehr als x-Millionen Sozialhilfeempfänger' ‚weniger als y-Prozent vom Bundeshaushalt für Kultur') zusätzlich von außen hinzukommen.

Installiert werden sie hauptsächlich über Kollektivsymbole wie z.B. ‚Netz' (‚das soziale Netz kann nicht mehr alle tragen') oder ‚Körper' (‚das Gesundheitswesen kollabiert'). Einzelne Kollektivsymbole können auch selbst flexibel nachgebessert werden: Galt das ‚soziale Netz' über Jahre hinweg wirklich als Sicherheitsnetz, in dem man bei sozialem ‚Absturz' aufgefangen wurde, so kursiert seit 1989 zunehmend die Formulierung vom ‚'Abfedern im sozialen Netz', was nichts anderes heißt, als dass der ‚Aufprall' jetzt unvermeidlich geworden ist, es nur noch um die Bandbreite des ‚Härtegrades' gehen kann.

Die Kollektivsymbole markieren aber nicht nur die Einschnitte auf den kontinuierlichen Kurven, sie zwingen das einzelne Individuum zudem förmlich, sich in einem ganz materiellen (körperlichen) Sinne dem flexibilisierten Gesellschaftskörper anzuschließen. Es genügt also, eine steigende Kurve (‚mehr Menschen werden älter') mit einer symbolischen Marke (‚1997 werden über zehn Millionen Bundesbürger über 80 Jahre alt sein') und einem Kollektivsymbol (‚eine nicht zu bewältigende Flut von Pflegefällen kommt auf uns zu') zu kombinieren, um den Effekt ‚es besteht dringender Handlungs- / Normalisierungsbedarf' zu erzielen und zudem die Individuen sich selbst, entsprechend der neuen Vorgabe ‚einstellen', zu lassen.

Diese Form des Normalismus wird von Link ‚flexibler Normalismus'genannt (im Gegensatz zu allen Formen von Protonormalismus, der auf der Basis einmal festgelegter Grenzen operiert). Kann die Festlegung fixer Grenzen durch ‚Dressur' und ‚Repression' erfolgen, so ist solche ‚Außenlenkung' mit dem Flexibilitäts-Normalismus unvereinbar. Damit er funktionieren kann, müssen die Subjekte imstande sein, sich selbst zu normalisieren. Diese Fähigkeit zur Selbst-Normalisierung setzt Innen-Lenkung voraus. Bekanntestes Beispiel für solche Selbst-Normalisierung, auf Grundlage von im Mediendiskurs verbreiteten Normalitätskurven, dürfte im Sog des Kinsey-Reports und ähnlicher populärer Studien wohl der gesamte Bereich der Sexualität (Beischlaffrequenzen, Aktivitätsaltersgrenzen, Anzahl der Partnerwechsel usw.) sein. Der lebensbegleitende, massenhafte Konsum des modernen medio-politischen Diskurses muss in diesem Sinne auch als eine immer wieder neu erfolgende, flexibel-normalistische ‚Selbst-Einstellung' der Subjekte verstanden und analysiert werden — und eben nicht als platte Manipulation. Kurz: Ohne die massenmediale Verbreitung von Daten, Kurven, Durchschnitts- und Grenzwerten, die immer auch ein Angebot zur Selbst-Normalisierung sind, wüsste niemand, was sozialpolitisch mehr oder weniger ‚normal' ist, und was jenseits der aktuell akzeptierten Grenzen liegt.

Wenn es zunächst einmal keine Wesensgrenze im statistischen Kontinuum gibt, dann können prinzipiell alle ‚anormal' sein. Daraus resultiert ‚Denormalisierungsangst' und für viele Politikbereiche Zustimmung zu jenen Optionen, die insgesamt eine $\frac{2}{3}$-Gesellschaft ausmachen. Pointiert gesprochen: Wer sich die Frage: ‚Bin ich normal? ' ernsthaft stellt, der ist unter den Bedingungen des Flexibilitätsnormalismus nahezu gezwungen, eine $\frac{2}{3}$-Gesellschaft zu akzeptieren. Denn es ist schwer, sich gegen eine Gesellschaft zu artikulieren, zu deren wichtigsten konstituierenden Mechanismen eben die normalistischen gehören.

4.5 Rückbindung des Exkurses

Wie schon dargestellt, beruht die Umsetzung des ‚clinical significance'-Ansatzes auf einer kontinuierlichen Größe und einer Grenze auf dieser (Normalität). Die Analogie zu den von Link aufgezeigten alltagssprachlichen Beispielen sollte deutlich sein. Es ist davon auszugehen, dass diese alltägliche, ‚normale' Logik hinter dem Rücken des Forschers diesen bestimmt (Normalismus als Interdiskurs).

Folgende Punkte schließen an die vorliegende Arbeit an:

- Im Gegensatz zur Norm, die eine binäre Entscheidung ermöglicht, ist Normalität eine kontinuierliche Größe. Verglichen mit dem CS-Konzept ist festzustellen, dass gerade hier versucht wird, via einer kontinuierlichen Größe eine binäre Entscheidung zu ermöglichen (vgl. insbesondere die in Abschnitt 2.2 auf Seite 7 diskutierte Problematik). Insgesamt bleibt der CS-Begriff der Logik der Normalität verhaftet.

- Die Tendenz des Individuums zur Selbstnormalisierung, d.h. die Antizipation der Stellung in der im Hintergrund schwebenden Verteilung aller Individuen, ist zwar nicht als solche relevant für die vorliegende Arbeit, könnte jedoch als Veränderungsmechanismus wirksam sein.

- Voraussetzung für die Normalisierung ist die Gleichartigkeit der Phänomene. D.h. von der Komplexität des Individuums muss abstrahiert werden. Lateinisch bedeutet abstractere abziehen. Dies verweist auf den Vorgang als ganzen: Eine Eigenschaft muss der anderen vorgezogen werden, nach dieser ausgewählten Eigenschaft wird erst die Reihe konstituiert. Natürlich ist es so, dass die Menge der nicht gewählten Eigenschaften für jedes Individuum eine andere ist, das bedeutet auch, dass für jedes Individuum diese ausgewählte Eigenschaft eine andere Bedeutung annimmt. Dieser Vorgang ist konstitutiv für eine statistisch-empirische Psychologie überhaupt (vgl. das Kapitel 5 auf Seite 42).

- Der erwähnte Zirkelschluss ‚Normal ist, was normalerweise als normal gilt' ist strukturgleich[1] dem in Abschnitt beschriebenen 2.2.2 auf Seite 11 Münchhausenprizip : Die Zugehörigkeit zu einer Gruppe (Verteilung der Kranken/Gesunden) definiert, die spätere Gruppenzugehörigkeit einer kontinuierlichen Größe (Maß).

- Auf den die diskursverbindene Funktion der Normalität ist ein Großteil des ‚Einleuchteffekts' der klinischen Signifikanz rückführbar.

- Bei dem in dieser Arbeit benutzten Normalitätsbegriff handelt es sich natürlich nicht um ein Beispiel des flexibeln Normalismus. Eine Essstörung ist eben eine Erkrankung und die Grenzen zu dieser sind nicht willkürlich verschiebbar: Ein niedriges Gewicht hat ganz handgreifliche körperliche Folgen, ein Verschieben der Grenzen ist nicht möglich. Es handelt sich um ein Beispiel des fixistischen

[1] Zum Begriff der Strukturgleichheit vgl. Carnap (1928, Seite 13).

> Protonormalismus. Ein Gegenargument wäre der in den letzten Jahren zu verzeichnende Trend zu einem geringeren Gewicht (Schönheitsideal), das heißt, die Grenze zur Essstörung wäre doch willkürlich. Dieses Gegenargument verliert jedoch seine Plausibilität eben wegen der körperlichen Konstanten.

Das eigentlich überraschende Ergebnis dieses Abschnittes ist, dass die Logik der Statistik nicht in einem abgesonderten Spezialdiskurs abgeschieden ist, sondern als Normalismus viele Diskurse miteinander verbindet. Unmittelbare Konsequenzen für die vorliegende Arbeit ergeben sich jedoch nicht, deshalb ist dieses Kapitel auch ein ‚Exkurs'. Jedoch erklärt dieser Exkurs einen Teil der Sinnhaftigkeit des CS-Kriteriums.

5 Exkurs: Das Wesen der Wissenschaft — Heideggers Wissenschaftstheorie

Bevor Heideggers Theorie der Wissenschaft dargestellt wird, soll zum einen diese Wahl (Warum gerade Heidegger?) durch die innere Logik dieser Theorie, zum anderen durch den Aufweis der Differenzen zu anderen Bestrebungen, die empirische Wissenschaft als solche theoretisch wiederzugeben, gerechtfertigt werden. Hinsichtlich der Bedeutsamkeit der vorliegenden Arbeit sei auf den Punkt 5.3 auf Seite 48 verwiesen.

Bevor jedoch Heideggers Diagnostik der bestehenden Wissenschaft wiedergegeben wird, müssen die Punkte kurz erläutert werden, die eine Einschränkung der Darstellbarkeit der Gedankengänge Heideggers bedingen. Heidegger bezeichnet seine späten Aufsätze oft als Wege durch die Sprache, d.h. es findet sich in den Texten sowohl eine gedankliche Entwicklung, als auch ein eigentümlicher Bezug zur Sprache. So betont Heidegger immer wieder, dass nicht auf Inhalte geachtet werden soll, sondern auf den Verlauf der Argumentation: „. . . auf den Weg zu achten, weniger auf den Inhalt. Beim Inhalt zu verweilen, verwehrt uns schon der Fortgang des Vortrages.“ (Heidegger, 1996, Seite 9). Aufgrund des partikulären Interesses im Rahmen dieser Arbeit sich nur auf die Theorie der Wissenschaft zu beschränken, wird also nicht der Weg (Fortgang) dargestellt, sondern gleichsam ein Ort, der an diesem Weg liegt.

An welchen Argumentationstellen finden sich also Gedanken zur Wissenschaft? Hier lassen sich vor allem zwei Punkte indentifizieren. Wissenschaftlichkeit (Technik) ist einmal Ausgangspunkt, zu dem Heidegger freilich immer wieder, wenn auch mit einem gewandelten Verständnis, zurückkehrt. Zum anderen kehrt das wissenschaftliche Denken in der Funktion des Zweifels am Heideggersschen Gedankengang wieder (wohl auch um die gängige Kritik an Heideggers Denken vorwegzunehmen): „Hat das Vorgebrachte noch das Geringste mit Wissenschaft zu tun? Es wird gut sein, wenn wir möglichst lange in solcher Abwehrhaltung ausharren. Denn so allein halten wir uns in dem nötigen Abstand für einen Anlauf, aus dem her vielleicht dem einen oder anderen der Sprung in das Denken des Bedenklichsten gelingt.“(Heidegger, 1990b, Seite 127). Das heißt, gerade der wissenschaftlich-technische Zweifel schafft eine Distanz zu den Heideggerschen Gedanken, welcher es wiederum ermöglicht den eigenen Ort (das Bedenklichste) näher zu betrachten (Rückkehr zum Ausgangspunkt). So ist die Theorie ebensosehr zutreffendes Abbild des wissenschaftlichen Fortschritts als auch Vorbereitung zu einem anderen Denken.

5.1 Warum gerade Heidegger?

Aus welchen Gründen ist die Lektüre von Heideggers Denken für uns und innerhalb einer eher mathematisch-methodischen Arbeit von Interesse?

Heidegger zählt sicher zu den bedeutendsten Denkern des letzten Jahrhunderts. Nach Heideggers eigener Theorie hat die Wahrheit in den Abschnitten der Geschichte immer eine andere Form. Sie besitzt einen Zeitkern. Der Denker entspricht dem Geschehen, dass allererst Wahrheit produziert, indem er passiv auf dieses (in Heideggers Sprache ‚Zuspruch') achtet (‚hört'). Oder noch konkreter: Der Denker schreibt den Gang des Gedankens, der der Logik des Sachverhalts entspricht, nur nieder: „Dem Anspruch der Weite, dem Anspruch des Verhaltens dieses Weltalters entsprechen wir, wenn wir beginnen, uns zu besinnen, indem wir uns auf den Weg einlassen, den jener Sachverhalt schon eingeschlagen hat, der sich uns im Wesen der Wissenschaft, jedoch nicht nur hier, zeigt."(Heidegger, 1990c, Seite 65). Die Größe des Denkers wäre nach Heidegger folglich durch das Maß zu bestimmen, inwieweit dieser das geschichtliche Wahrheitsgeschehen in größtmöglicher Klarheit niedergeschrieben hat. Nach diesem Maßstab könnte also Heidegger als bedeutender Denker bezeichnet werden. D.h. ein Grund die Wissenschaftstheorie Heideggers darzustellen ist deren Klarheit. Weiterhin enthält seine Diagnostik in durchdachter Form Elemente verwandter Positionen.

Demgegenüber wären aus Sicht Heideggers die anderen Wissenschafttheorien ebenfalls bestimmt durch das geschichtliche Wahrheitsgeschehen, einmal als dessen Symptom, ein anderes Mal als dessen teilweises Abbild. Kurz zu nennen wären:

- Die *konstruktivistische Position* legt den Schwerpunkt ihrer Theorie auf den schöperischen Akt, der mit jeder Empirie verbunden ist (Glasersfeld, 1997; Schmidt, 1987): Der Mensch findet weniger die Gegenstände vor, als dass er sie erfindet. Die Dinge sind hierbei immer schon durch eine auf Zwecke ausgerichtete Methodik für uns gegeben, d.h. es besteht kein unmittelbarer Zugang zu den Dingen, sondern nur einer durch unsere Sinne (oder technischen Hilfmittel) und Weltverständnis vermittelter. Resultat ist folglich ein relatives Wissen. Die Dinge sind das Ergebnis einer Anwendung einer bestimmten Methodik (Konstruktion). Dieser Gedankengang ist sehr Nahe der Heideggerschen Diagnostik des technischwissenschaftlichen Prozesses. Jedoch ist dies für Heidegger nur eine Wirkung des Wahrheitsgeschehens, das in unserer Epoche vorherrscht. Das bedeutet, im Gegensatz zum Konstruktivismus beharrt Heidegger auf der Möglichkeit, dass die Dinge sich uns auf anderer Weise zeigen können (wie dies schon in vorhergehenden Epochen der Fall war).

- Kuhns Theorie der wissenschaftlichen Revolutionen (Kuhn, 1967) fokussiert auf der Abfolge herrschender Paradigmen, die das Forschungsfeld strukturieren in dessen Licht sich die Empirie zeigt. Ein Paradigma wird von einem anderen Paradigma abgelöst (wissenschaftliche Revolution), wenn sich die Beobachtungen häufen[1], die mit dem alten Paradigma nicht mehr vereinbar sind. Hier besteht

[1]Hier besteht eine Parallele zu Piagets Therorie der menschlichen Entwicklung. Auch hier wird eine

nach Heidegger eine Analogie bzgl. des kategorialen Charakter des geschichtlichen Prozesses.Im Unterschied zu Kuhns Theorie greift bei Heidegger die Unterscheidung, die Epochen voneinander trennt, tiefer: sie betrifft das Ereignis Wahrheit selbst.

- Die kritische Theorie[1] betont die gesellschaftliche und historische Bedingtheit des Wissenschaftsprozesses. Das bedeutet, dass auch die Empirie nicht als unmittelbare Gewissheit vorhanden ist, sondern jedes empirisches Faktum begrifflich vermittelt ist. Letztendlich findet sich also, wie bei Heidegger, ein ‚Zeitkern' der wissenschaftlichen Wahrheit.

Welchen Zweck hat die Beschäftigung mit der Wissenschaft *für Heidegger* selbst? Hier ist, um Missverständnissen vorzubeugen, vorweg die Identifikation der Wissenschaft mit der Technik zu erwähnen. D.h. die oben genannte Frage richtet sich sowohl an die Technik als auch an die Wissenschaft. Zum Einen soll sie uns in die Lage versetzen zu der Technik eine „freie Beziehung zu ihr vorzubereiten"(Heidegger, 1990a, Seite 9), denn erst wenn die Technik als ihr Wesen (oder währen, d.h. was sie für uns ist, ihre Bedeutsamkeit) offenbar wird, kann der Mensch in ein freies Verhältnis zu dieser gelangen. Denn dieses ist weder frei wenn die Technik bejaht oder verneint wird noch wenn eine neutrale Position zu ihr eingenommen wird.

5.2 Heideggers Diagnostik des Wissenschaftsgeschehens

Dem mittelalterlichen Blick mag eine moderne, psychologische Arbeit suspekt und unverständlich erscheinen — ähnlich vielleicht unserem Erstaunen, wenn wir von der Ernsthaftigkeit von mittelalterlichen Untersuchungen hören, die sich etwa um die Frage bemühten, welcher Sprache sich Gott bediente, um sich mit Adam und Eva zu verständigen. Hier ist vielleicht schon der historische Bruch spürbar, der die verschiedenen Epochen mit ihrer jeweiligen Wissenschaftlichkeit radikal trennt. Wissenschaft ist eben kein linearer Fortschritt, also ein *mehr* und *besser* dessen was früher war, sondern selbst Auswirkung unterschiedlicher, die Epochen bestimmende Faktoren. Heidegger benennt dies, vielleicht überraschend, als ‚Metaphysik' eines Zeitalters: „Die Metaphysik begründet ein Zeitalter, indem sie ihm durch eine bestimmte Auslegung des Seienden und eine bestimmte Auffassung der Wahrheit den Grund seiner Wesengestalt gibt" (Heidegger, 1980, Seite 73). Dieser Grund bestimmt nicht nur das, was die Dinge *sind*, es bestimmt auch das, was für ein Zeitalter gegenüber den anderen das Auszeichnende ist, in unserem gegenwärtigen Zeitalter vor allem auch die Wissenschaft und Technik.

Den radikalen Unterschied der Epochen verdeutlicht Heidegger am Beispiel der Physik: „So kann man auch nicht sagen, die Galileische Lehre vom freien Fall der Körper sei wahr und die des Aristoteles, der lehrt, die leichten Körper streben nach oben, sei

Entwicklungsstufe durch eine andere abgelöst, wenn sich Beobachtungen häufen, die nicht mehr in das herrschende Schemata einbaubar (assimilierbar) sind (Piaget & Inhelder, 1986).

[1] Sicher gibt es nicht die kritische Theorie als solche, sondern nur bestimmte Vertreter mit unterschiedlichen Positionen. Betrachtet werden hier nur einige gemeinsame Punkte.

falsch, denn die griechische Auffassung vom Wesen des Körpers und des Ortes und des Verhältnisses beider ruht auf einer anderen Auslegung des Seienden und bedingt daher eine entsprechend verschiedene Art des Sehens und Befragens der Naturvorgänge“ (Heidegger, 1980, Seite 75). D.h. unsere Art und Weise zur Wahrheit zu kommen läßt sich nicht auf andere Epochen übertragen, will man diese nicht grundlegend missverstehen. Die *neuzeitliche Naturwissenschaft* unterscheidet sich sowohl von der mittelalterlichen ‚doctrina‘ und ‚scientia‘ als auch von der alten griechischen Form der Wissenschaft (ἐπιστήμη). Abbildung 5.2 gibt einen Überblick zur Theorie und der sich ergebenden Wirklichkeit der Gegenstände in den verschiedenen geschichtlichen Stadien.

Epoche	**Theorie**	**Wirklichkeit**
Griechische Wissenschaft	Theorie (θεορία) bedeutet „den Anblick, worin das Anwesen erscheint, ansehen und durch solche Sicht bei ihm sehend verweilen“. Wahrheit ist Unverborgenheit (ἀλέδεια), Theorie ist „das hütende Schauen der Wahrheit“ (im Sinne der alten Bedeutung von hüten=wahren).	Die Dinge sind in das Anwesen her – (ins Unverborgene) – vor (ins Anwesen) gebrachte. Entweder durch sich selbst (Natur, ψύσιζ) oder durch den Menschen. Die Dinge sind im Modus des sich-in-der-Vollendung haltens (ἐντέλεχεια).
Römische Wissenschaft	Aus der griechischen θεορία wird die lateinische Contemplatio. Das bedeutet: „ewas in Abschnitte einteilen und darin umzäunen“. D.h. der „Charakter des eingeteilten, eingreifenden Vorgehens gegen das, was ins Auge gefasst werden soll, macht sich im Erkennen geltend“.	Die Römer sehen die Wirklichkeit als vom actus bestimmt, d.h. als das, was die operatio ergibt. Die Wirklichkeit ist somit das Erfolgte, d.h. wird von einer Ursache bedingt (Kausalität).
Neuzeit	Hier erscheint die Wissenschaft in der Form der Betrachtung, wobei trachten im Sinne von verfolgen, nachstellen um es „sicher zu stellen“ (Methode) zu verstehen ist, d.h. Theorie ist jetzt das „nachstellende und sicherstellende Bearbeiten des Wirklichen“. Sie „stellt jeweils einen Bezirk des Wirklichen als ihr Gegenstandsgebiet sicher“.	Die Dinge sind jetzt das Tatsächliche, was durch die Wissenschaft sichergestellt werden kann (im Experiment). Die Dinge zeigen sich jetzt im Modus des Erfolgens: „Der Erfolg ergibt, dass das Anwesende durch ihn zu einem gesicherten Stand gekommen ist und als solcher Stand begegnet“ (Gegenstand).

Tabelle 5.2: Theorie und Wirklichkeit in den verschiedenen wissenschaftlichen Epochen (Heidegger, 1990c)

Wodurch ist nun die griechische ἐπιστήμη ausgezeichnet? Diese ist nicht so sehr durch ein technisches Taxieren der Dinge gekennzeichnet sondern durch ein „[...] den Anblick, worin das Anwesende erscheint, ansehen und durch solche Sicht bei ihm sehend verweilen“(Heidegger, 1990c, Seite 48). D.h. die Dinge werden abseits einer Nutzbarmachung rein geschaut, sind das, als was sie sich zeigen. Die Dinge sind vom Verboge-

nen ins Unverborgene, in das Anwesen, gebracht worden.

Die *römische Wissenschaft* übersetzt die griechische Schau (Theorie, θεορία) als Contemplatio. Übersetzt heißt dies: in Abschnitte einteilen und umgrenzen. D.h. mit diesem Gewicht auf das analytische Zerlegen hat sich die römischen Wissenschaft schon erheblich von der griechischen entfernt. Die Dinge zeigen sich jetzt in einem bestimmten Verständnis der Kausalität, sie sind durch eine Ursache bedingt.

Dagegen läßt die *mittelalterliche scientia* die Dinge nicht von sich her sich zeigen sondern fokussiert auf das maßgebende göttliche Wort: der Bibel. Somit ist die Wissenschaft des Mittelalters entweder Auslegung des göttlichen Wortes (Hermeneutik) oder das Wiederfinden dessen in der Schöpfung.

Neuzeitliche Naturwissenschaft ist dagegen ausgezeichnet durch Exaktheit und Gesetz, Forschung, Experiment und Mathematik, durch Institutionalisierung und Betrieb, durch Technik, Spezialisierung und System. Wie stellt sich nun nach Heidegger unser ‚Wissenschaftbetrieb' dar?

Abstrakt formuliert ist bei dem späten Heidegger das Wissenschaftsgeschehen als ein in sich doppelter Vorgang (Heidegger, 1980, Seite 81f.). Dieser besteht einmal in dem Hineintragen bestimmter Strukturen und durch dieses Hineintragen konstituiert sich das wissenschaftliche Feld. Es wird ein Grundriss des Gegenstandsbereiches entworfen. So stellen etwa folgende Sätze einen Teil des Grundrisses der Physik dar: „Keine Bewegung oder Bewegungsrichtung ist vor der anderen ausgezeichnet. Jeder Ort ist jedem anderen gleich[...]" Und erst durch das Hineinsehen des Gegenstandes in diesen Grundriß wird dieser als Naturvorgang sichtbar(Heidegger, 1980, Seite 77). Alle Erkenntnisse werden jeweils auf diesen Entwurf zurückbezogen und nur dasjenige ist Gegenstand der Wissenschaft, das in den Entwurf passt. Mit diesem Prozeß verbunden ist also die Konstituierung als (Fach-) Wissenschaft und ermöglicht, in diesem, die Bewegungen der Wissenschaft, also die Art des Vorgehens, mithin die Methodik. Eben durch den Entwurf des Grundrisses kommt der Vorrang der Methode über den Gegenstand zustande. Diese ist nun für den Wissenschaftler bindend. Die Bindung der Wissenschaften als Einrichtung von Zusammenhängen von ist die ihnen zugehörige Strenge. Die Entfaltung der Strenge vollzieht sich in den Weisen des Vorgehens (Methodik). Dieses Vorgehen bringt den Gegenstandsbezirk jeweils in eine bestimmte Richtung der Erklärbarkeit. Als ausschließend gegenüber anderen möglichen Vorgehensweisen nimmt die Bindung an den Entwurf gewaltsamen Charakter an. Seine innere Grenze ist das Ausschließen anderer Weisen in der Form des ‚Nicht'.

Ein Hauptkennzeichen neuzeitlicher Wissenschaft ist das Experiment. Auch die griechischen und mittelalterlichen Forscher beobachteten die Dinge (ἐμπερεία=Empirie, lat: experienta) ihre Eigenschaften und Veränderungen unter wechselnden Bedingungen und erbrachten also ein Wissen um die Weise, wie sich die Dinge in der Regel verhalten (Heidegger, 1980, Seite 78f.). In der neuzeitlichen Wissenschaft wird das Experiment unter den Bedingungen eines Gesetzes realisiert: „Ein Experiment ansetzen heißt: eine Bedingung vorstellen, dergemäß ein bestimmter Bewegungszusammenhang in der Notwendigkeit seines Ablaufes verfolgbar und daher für die Berechnung im voraus beherrschbar gemacht werden kann". Dieses Experiment wird auf dem Hintergrund des Grundrisses des Gegenstandsbezirks vollzogen: „Dieser gibt das Maß und bindet

das vorgreifende Vorstellen der Bedingung“. Die Erklärung des Experimentes „[...] begründet ein Unbekanntes durch ein Bekanntes und bewährt zugleich dieses Bekannte durch das Unbekannte“.

Die Bindung der Wissenschaft ist in heutiger Zeit eine technische, genauer: eine mathematische. Das im voraus schon Bekannte nannten die Griechen das Mathematische Τὰ ματέματα. Dabei ist das, was wir unter Mathematik verstehen nur ein Sonderfall dieses Begriffs. Warum ist diese Bedeutungsverschiebung zustande gekommen? Da die Zahlen, also weiter benannt ist das Quantitative, ‚das Aufdringlichste‘, vorher bekannte sind. Weniger aufdringlich, jedoch nicht weniger ausschlaggebend, ist das vorhergehende Wissen über die Gegenstände der Wissenschaft. So ‚weiss‘ der Physiker im voraus, das keine Bewegung einer anderen vorzuziehen sei, seine Gegenstände eine Trägheit besitzen etc. Dieses letztgenannte variiert natürlich sowohl von Wissenschaft zu Wissenschaft (wenn in der Psychologie von der Trägheit des Menschen die Rede ist, ist wohl anderes gemeint), grundsätzlich ist es mithin gegenstandskonstituierend. In Heideggers Worten: ein bestimmter Grundriss der Naturvorgänge wird entworfen. Dabei ist beim Wort ‚Entwurf‘ auch das ganz handfeste als ‚etwas hineinwerfen‘ mitzuhören. Der Entwurf *zeichnet* vor, in welcher Weise das erkennende Verfahren sich an den eröffneten Bezirk zu binden hat.‘ Die alte Bedeutung von Reißen war ‚schreiben‘, ‚zeichnen‘, eigentlich ‚ritzen ‘ (Kluge, 1999). Es handelt sich also um einen Schriftcharakter. In diesen Grundriss eingetragen ist, dasjenige, was für das Erkennen künftig sein Gegenstand sein soll. Der Gegenstand wird, jetzt über Heidegger hinausgehend, abgelesen. Wobei die Art des Schriftcharakters die Art des Lesens bestimmt. Eben aus dem Hineinsehen der Gegenstände in den Grundriss ergibt sich die ‚Strenge‘ der Wissenschaften. Durch die Messung „mit Hilfe der Zahl und der Rechnung“ (Heidegger, 1980, Seite 77) ergibt sich der exakte Charakter der Naturwissenschaft.

Die Wissenschaft ist nun eine Einrichtung des eines Wissens (des Entwurfs). Hier kommt die Unterscheidung von wahr und richtig ins Spiel: Innerhalb der Wissenschaft gibt es nur ein richtig oder falsch. Der Vorgang jedoch, der die Wissenschaft Wissenschaft sein lässt, ist ein Wahrheitsgeschehen. Dieses Wahrheitsgeschehen, besser: Entbergung lässt die Dinge aus dem Verborgenenen in das Unverborgene kommen. D.h. die Dinge sind je schon durch eine Entbergung das, als was sie uns erscheinen. Eine bestimmte Art der Entbergung nun bringt uns vor die Möglichkeit der Entscheidung, ob ein Sachverhalt richtig oder falsch ist. Wesentliches Kennzeichen unserer Wissenschaft ist, dass sie *technisch* ist. Es zeigt sich, „dass die Wissenschaft im Weltkreis des Abendlandes und in den Zeitaltern der Geschichte eine sonst nirgends auf der Erde antreffbare Macht entfaltet hat und dabei ist, diese Macht schließlich über den ganzen Erdball zu leben“. Als eigenständige Macht unterliegt sie nicht dem Willen des Menschen, so kann sie etwa nicht durch Beschlüsse abgeschafft werden. Vielmehr waltet hier ein „größeres Geschick“. Was ‚waltet‘ hier? „Das in der modernen Technik waltende Entbergen ist ein Herausfordern, das an die Natur das Ansinnen stellt, Energie zu liefern, die als solche herausgefördert und gespeichert werden kann“. Hier zeigt sich der Charakter der Technik: Sie ist ein ‚stellen‘, insgesamt bezeichnet Heidegger die Gesamtheit des Stellens mit dem Kunstwort „Gestell“. Das Gestell, ein substantiviertes Verb, ist nicht Instrument des Menschen, sondern der Mensch ist selbst in dieses hin-

eingestellt. Prägnantes Beispiel ist hier die heutige Gentechnik. Hier ist der Mensch als Ganzes von der Technik in einen Machbarkeitszusammenhang gestellt, der Mensch ist in diesem ein Gegenstand technischer Überlegungen. „Die Natur ist zum Gegenstand und zwar eines Vorstellens, das ihre Vorgänge als berechenbaren Bestand herausstellt und sichert“ (Heidegger, 1997, Seite 100)

Jedoch sind die Wissenschaften an sich nicht vollständig. Sie benötigen jeweils ein Unumgängliches, was sie selbst hält, jedoch mit ihren eigenen Mitteln nicht erreichbar ist: „Natur, Mensch, Geschichte, Sprache bleiben für die genannten Wissenschaften das innerhalb ihrer Gegenständigkeit schon waltende Unumgängliche, worauf sie jeweils angewiesen sind, was sie jedoch in seiner Wesensfülle durch ihr Vorstellen jedoch nie umstellen können“ (Heidegger, 1990c, Seite 60). Für die Physik ist das die Natur, für die Psychiatrie (und um Heidegger zu paraphrasieren auch die Psychologie) das Dasein des Menschen. Das Unumgängliche ist das Wesen des Gegenstandes der Wissenschaft. Daraus ergibt sich der paradoxe Sachverhalt, dass die Wissenschaft dieses Unumgängliche benötigt um sich als Wissenschaft einzurichten, dieses selbst mit ihren Mitteln jedoch nicht erreichen kann[1]. Unumgänglich ist dies Konstituenz der Wissenschaft aus zwei Gründen: Einmal wegen der Angewiesenheit auf das Anwesen des wissenschaftlichen Gegenstandes, zum anderen da sich die Wesensfülle des Gegenstandes nicht der Methode erschließt. Für die Wissenschaft ist dieses Unumgängliche nun selber unscheinbar: „Aber er liegt nicht in Ihnen wie der Apfel im Korb. Wir müssen eher sagen: Die Wissenschaften ruhen ihrerseits im unscheinbaren Sachverhalt wie der Fluß im Quell“. Hier kommt Heideggers eigener Vorschlag zu Analyse der Wissenschaftlichkeit ins Spiel: Diesem Unscheinbaren nachzusinnen (Besinnung): „Die Besinnung bringt uns dagegen erst auf den Weg zu dem Ort unseres Aufenthalts“.

5.3 Rückbindung des Exkurses

Bindet man diese Ausschnitte aus Heideggers Überlegungen zur Wissenschaft zurück in die vorliegende psychologische Arbeit, ist vor allem die Stellung des Menschen hervorzuheben: der Mensch ist in die naturwissenschaftliche Psychologie ‚gestellt‘ ist also auch eine Wirkung des ‚Gestells‘, qua Konstituens der Realität, darstellt. In diesem ist er ein Datenpunkt einer Verteilung. Dieser abstrakte Charakter der Wissenschaft wurde schon im Kapitel zur Normalität (siehe Abschnitt 4.5 auf Seite 40) herausgehoben und erscheint hier nochmals in einem allgemeineren Zusammenhang. Viel bedeutsamer für die vorliegende Arbeit ist jedoch die eigentliche, prozessuale Bestimmung des Wissenschaftgeschehens. Auch in der vorliegenden Arbeit geht es um Strukturen, die einen Gegenstand ‚produzieren‘. So ist zwar der ‚clinical significance‘-Ansatz sicherlich nichts radikal anderes als andere Erfolgsmaße, jedoch ebensosehr Struktur und hinein-

[1] Hier findet sich ein bemerkenswerte Parallele zu Žižeks Analyse der heutigen Warenform. Auch hier findet sich innerhalb der Ware etwas, was mehr ist als sie, was aber nie Teil der Ware sein kann: „Die materielle (‚reale‘) Leere im Zentrum steht natürlich für die strukturelle (‚formale‘) Lücke, aufgrund der kein Produkt ‚wirklich genau das ist‘, das heißt, kein Produkt erfüllt wirklich die Erwartung, die es weckt“ (Žižek, 2003, Seite 146)

getragen (‚angewandt') in ein Gegenstandsgebiet, dass schon von den Grundriss ‚naturwissenschaftliche Psychologie' erschlossen ist. Mit Heidegger sei hier insbesondere auch auf die strukturgebende Rolle der Mathematik verwiesen. Eben weil die Psychologie naturwissenschaftlich und damit exakt ist, ist sie mathematisch, rechnend.

6 Ein Vorschlag zur Modifikation des RC-Begriffs und die daraus resultierende Fragestellung der Arbeit

Ausgangspunkt der bisherigen Darstellung (siehe Kapitel 2 auf Seite 5) war der Begriff der Heilung. Dieser wurde in eine mathematische Begrifflichkeit übersetzt. Durch diesen Übersetzungsvorgang erhielt das Konzept neue Bestimmungen. Abbildung 6.1 visualisiert noch einmal diesen Weg.

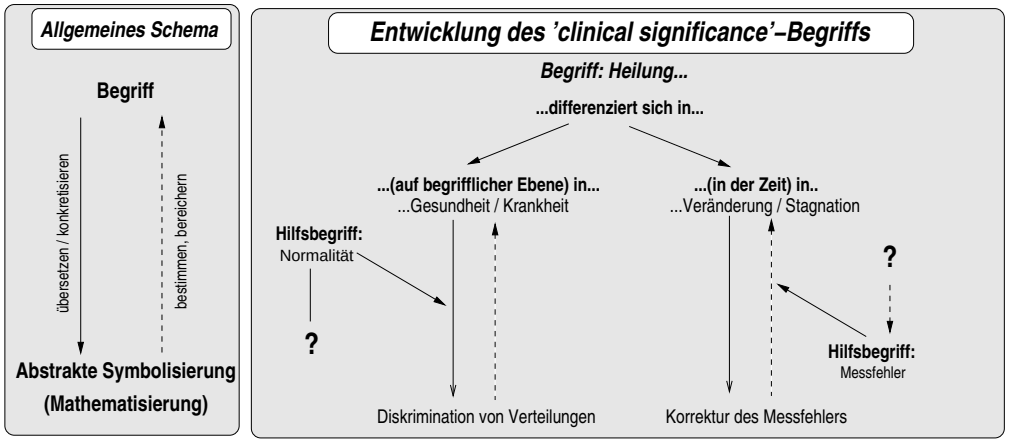

Abbildung 6.1: Schema zur Umsetzung des Heilungsbegriffs in eine mathematische Begrifflichkeit.

Die linke Seite der Abbildung 6.1 zeigt den Prozess auf einem abstrakten Niveau, während die rechte Seite die konkreten Schritte der Umsetzung darstellt. Auf der linken Seite der Abbildung 6.1 ist die mathematische Umsetzung des Heilungsbegriffs wiedergegeben. Der Begriff Heilung zergliedert sich einmal in den Zustandsbegriff Gesundheit mit dem ihm inhärenten Zuständen gesund/krank, das andere Mal, zeitlich, in dem Veränderungsbegriff und die zugehörigen Veränderungen Erkrankung / Gesundung. Benannt sind beide, jetzt neu entstandenen Unterkonzepte, als ‚clinical significance' und den ‚reliable change', welche den statischen und dynamischen Aspekt

des Heilungsbegriffes repräsentieren. Um diese beiden Konzepte in eine mathematische Formelsprache zu übersetzen, werden verschiedene Hilfsbegriffe, welche sich aus der statistischen Theorie ergeben, benötigt. Dies ist einmal hinsichtlich der mathematischen Diskrimination von zwei Verteilungen der Normalitätsbegriff, auf der anderen Seite hinsichtlich der Zuverlässigkeit der Veränderung der Reliabilitätbegriff. Es zeigt sich also, dass durch den Prozess der Mathematisierung ein Zwang entsteht, in diese andere Sprache andere Konzepte hinzuzuziehen oder Unterschiede einzuführen, die in der ursprünglichen Begrifflichkeit nicht vorkamen. Durch diese, aus der Logik der empirischen Forschung resultierenden Notwendigkeit kommt es hinsichtlich der Begrifflichkeit zu einer Bereicherung und Konkretisierung. Diese Momente des Prozesses sind auf der abstrakteren Ebene auf der linken Teilabbildung 6.1 wiedergegeben: Die Begrifflichkeit wird in die mathematische Formelsprache übersetzt, wodurch die auf diesem Weg durchgeführten Konkretisierungen und Abstraktionen wiederum die Begrifflichkeit bereichern und präzisieren. Also entsteht durch diesen Übersetzungs-Rückübersetzungprozess eine neue Bedeutung.

Die Darstellung und Begündung des *modifizierten* RC-Ansatzes kann ebenfalls auf beiden Ebenen, der eher begrifflichen als auch auf der eher mathematischen, erfolgen.

6.1 Begründung der Modifikation des RC-Ansatzes auf begrifflicher Ebene

Die Begründung auf einer begrifflichen Ebene kann aufgrund der in Abbildung 6.1 auf der vorherigen Seite erscheinenden Asymmetrie erfolgen: Während der Hilfsbegriff Normalität eher der begrifflichen Ebene zugeordnet werden muss, fehlt auf dieser Ebene ein Bedeutungsäquivalent für den Reliable-change Teil des Indizes. Dieser rührt von einem mathematischen Begriff her, dem der Wahrscheinlichkeit. Es könnte hier eingewandt werden, dass der mathematische Begriff der Reliabilität eine Übersetzung des alltagssprachlichen Begriffs der Zuverlässigkeit ist, also das Argument des fehlenden Bedeutungsäquivalentes gegenstandslos ist. Hier entsteht jedoch ein Widerspruch, dass die Zuverlässigkeit nur für einen Teil (RC), nicht jedoch für das ganze Konzept gelten soll. Würde diese nun für das ganze Modell angewandt werden, entstünde auf der Seite der Veränderung wiederum eine Leerstelle[1]. Inhaltlicher Fokus der Arbeit ist vielmehr die viel augenscheinlichere Asymmetrie des RC- Begriffs. Es gilt, diesen Begriff mit Bedeutung zu versehen. Was liegt nun näher als per Analogschluss zur CS die Normalität von Veränderungen zu fordern? Ein ähnliches Konzept findet sich jedoch, allerdings unter einem anderen Namen, bereits in der Psychotherapieforschung wieder.

[1]Eigentlich müsste also auch die Zugehörigkeit zu den Gruppen (linke Seite der Abbildung 6.1 auf der vorherigen Seite) als Wahrscheinlichkeit ausgedrückt werden (vgl. Abschnitt 2.2.3). Das soll heißen: Es ist keineswegs *sicher*, dass sich jemand in der Gesunden- und Kranken-Gruppe befindet, es ist nur mehr oder weniger *wahrscheinlich*. Konsequent wäre hier auf der mathematischen Ebene eine Umsetzung dieses Sachverhalts durch die Latent-class Analyse (Formann, 1984). D.h. auch auf der Ebene der ‚clinical significance' besteht eine Asymmetrie, die jedoch in der vorliegenden Arbeit weiter nicht behandelt wird.

Die bekannten, bzgl. der Effektivität von Psychotherapie getroffenen Schlussfolgerungen, die Eysenck (1957) durch den Vergleich psychoanalytischer und eklektischer Behandlungen neurotischer Patienten mit der base-rate unbehandelter Patienten aus Allgemeinkrankenhäusern zieht, war für die Psychotherapieforschung gleichermaßen schockierend, wie Motivation, diese empirisch zu widerlegen:

> „In general, certain conclusions are possible from these data. They fail to prove that psychotherapy, Freudian or otherwise, facilitates the recovery of neurotic patients. They show that roughly two-thirds of a group of neurotic patients will recover or improve to a marked extent within about two years of the onset of their illness, whether they are treated by means of psychotherapy or not. This figure appears to be remarkably stable from one investigation to another, regardless of type of patient treated, standard of recovery employed, or method of therapy used. From the point of view of the neurotic, these figures are encouraging; from the point of view of the psychotherapist, they can hardly be called very favorable to his claims.“ (Eysenck, 1957)

Eysenck (1957) vergleicht also zwei Bedingungen, unter denen Veränderungen stattgefunden haben, stellt die Gleichheit beider hinsichtlich der Effektivität fest und schließt auf die Wertlosigkeit der Psychoanalyse. Referenz dieses Vergleichs ist die sogenannte base-rate, oder die Spontanheilungsquote. Die methodische Umsetzung dieses Gedankens ist das randomisierte Kontrollgruppendesign, der, jedenfalls nach der evidence-based medicine so benannte Goldstandard. Verglichen wird also mit der Spontanheilungsquote. Im Grunde handelt es sich hier ebenfalls um den Vergleich mit etwas, das als Normalität der Veränderung bezeichnet werden kann, allerdings handelt es sich um die Normalität der Veränderung einer bestimmten Gruppe, der der unbehandelten Erkrankten (ein eingehender Vergleich des randomisierten Kontrollgruppendesigns mit der hier vorgeschlagenen Methodik erfolgt in der Diskussion).

6.2 Begründung der Modifikation des RC auf einer methodischen Ebene

Diese Normalität der Veränderung soll also in den bisherigen Ansatz der clinical significance integriert werden. Dies führt unmittelbar zu der Frage, wie dies zu geschehen habe, sowie zu den bisherigen Eigenschaften der RC-Formel beim Auftreten von wirklichen Veränderung. Anzumerken bleibt, dass die wirkliche Änderung von Eigenschaften nicht im Konzept der Retestreliabilität der klassischen Testtheorie enthalten ist. Diese geht nämlich von einem konstanten Merkmal (trait) aus. Wirkliche Änderungen werden natürlich in der klassischen Testtheorie berücksichtigt, hinsichtlich der Reliabilität stellen sie jedoch eine zu eliminierende Störgröße dar.

Reliabilität wird in der klassischen Testtheorie durch die Angabe von Korrelationen operationalisiert. Wie verhält sich diese Operationalisierung beim Auftreten unabhängiger Schwankungen? Abbildung 6.2 auf der nächsten Seite versucht dies durch eine

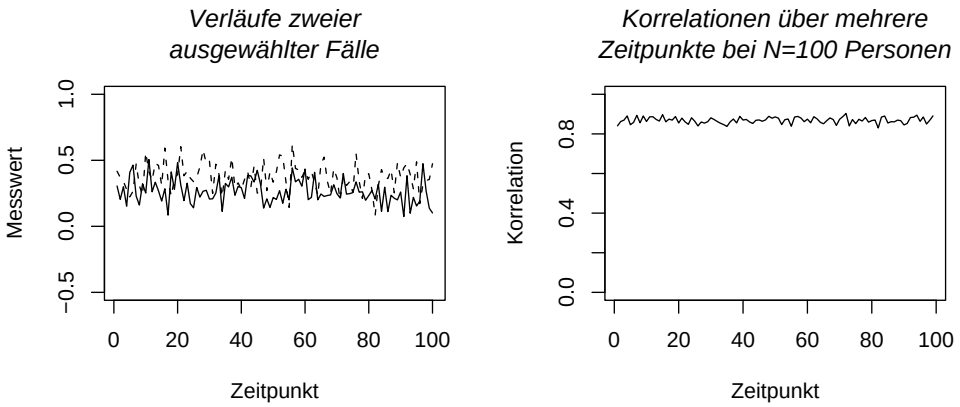

Abbildung 6.2: Das Veränderungsmodell der klassischen Testtheorie und die Konsequenzen für verschiedene Retestkorrelationen.

Simulation[1] zu illustrieren. In beiden Teilabbildungen sind auf der Abszisse 100 Messpunkte realisiert. Die Ordinate der linken Teilabbildung zeigt die Merkmalsausprägung zweier Fälle pro Messzeitpunkt. Die rechte Teilabbildung zeigt hingegen die Korrelationen des n-ten mit dem ersten Messzeitpunktes. Hierbei wurde eine Stichprobe von 100 Personen zur Simulation verwandt. Wie im linken Teil der Grafik ersichtlich, fluktuiert der Messwert der beiden ausgewählten Personen um deren personenspezifischen wahren Wert. Diese Schwankungen sind zufällig, es findet sich also keine Tendenz über die Zeit, wie etwa eine kontinuierliche Zunahme oder Abnahme. Im Kontext der klinischen Psychologie könnte man also von einem Fehlen von Erkrankungs- (Zunahme) oder Gesundungsprozessen (Abnahme) sprechen. Die Personen haben sozusagen *ihre* Merkmalsausprägung, die von Messzeitpunkt zu Messzeitpunkt mehr oder weniger schwankt. Das zeigt sich auch in den beobachteten Retestkorrelationen. Diese sind von Messzeitpunkt zu Messzeitpunkt mehr oder weniger gleich, auch sie schwanken um einen bestimmten Wert (rechter Teil der Abbildung). Das Ausmaß der individuellen Schwankungen beeinflusst die Korrelationen vor allem in ihrer Höhe. Wären die beiden dargestellten Personen also variabler in ihrer Merkmalsausprägung, läge die Zickzacklinie der Korrelationen auf einer anderen Ebene, nämlich unter der dargestellten Linie, d.h. die Retestreliabilität wäre geringer. Wie bekannt, kann man durch diesen Verlauf der Individuen eine Regressionslinie ziehen. Diese hätte eine Steigung von Null. Würde man sich jetzt die Verteilung dieser Veränderungswerte vorstellen, so hätten alle Personen den gleichen Wert von Null.

Lässt man jedoch auch Personen in der Stichprobe zu, die sich wirklich verändern, die also etwa erkrankten oder gesunden, so sähe die Verteilung der Veränderung an-

[1] Die Simulationsvorgaben waren wie folgt: Es wurde zu jedem Messzeitpunkt pro Person per Zufall ein personenspezifischer Wert simuliert $N(T_P, \sigma_p)$, wobei der wahre Wert der T_P zeitunabängig ist und die Fehler nicht korreliert sind. Es handelt sich um die Grundaxiome der klassischen Testtheorie.

ders aus.

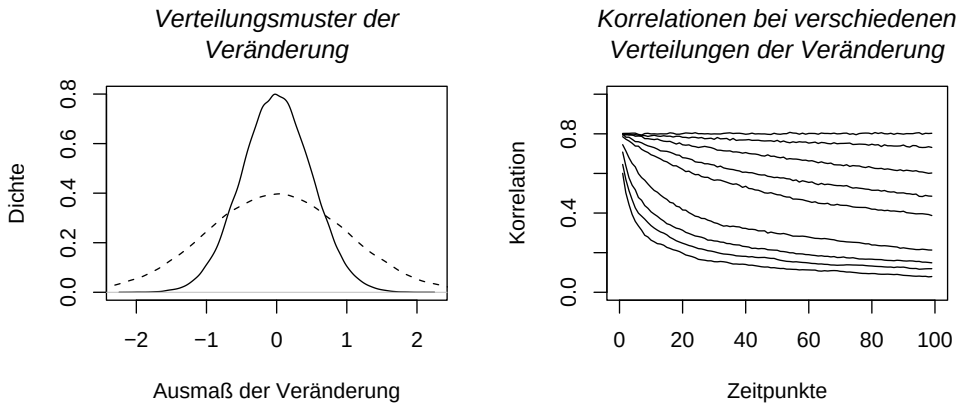

Abbildung 6.3: Verschiedene Verteilungen von Veränderungen und die daraus folgenden Retestkorrelationen.

Die linke Teilabbildung von Abb. 6.3) zeigt zwei Verteilungen von Veränderungen[1]. Dargestellt wird die Verteilung der individuellen Regressionslinien (slopes, eigentlich der Steigung derselben). D.h. jeder Punkt dieser Verteilung repräsentiert eine Steigungskurve eines Individuums, als Gruppe bilden sie eine Verteilung. In der schmalen Kurve ist der Anteil an Personen, die sich weniger verändern, größer als in der weiteren Verteilung. Das heißt, beide Verteilungen unterscheiden sich vor allem in ihrer Variabilität. Veranschaulicht werden kann dies vielleicht durch den Vergleich der Veränderungen einer Skala, welche eine eher schwere klinische Störung abbildet mit einer Skala, welche eher eine leichtere abbildet. So ist etwa denkbar, dass sich in der Normalbevölkerung relativ wenig Menschen finden, die sich hinsichtlich paranoider Vorstellungsinhalte verändern, dass also die Verteilung der Veränderungswerte relativ schmal ausfallen wird. Dagegen mag ein Merkmal wie Depressivität ein größeres Ausmaß an Veränderungen zeigen, d.h. die Verteilung wird breiter sein. Allerdings kann bei beiden Skalen angenommen werden, dass sich der Durchschnittswert (eigentlich der zu erwartende Wert) in der Bevölkerung nicht ändert. Dies mag bei essstörungsrelevanten Skalen anders sein. Folgt man etwa der Argumentation, dass diese kulturell bedingt sind, ist etwa mit einer Erhöhung/Verminderung des Durchschnittswertes durch gesellschaftliche Trends zu rechnen. Wie würde sich dies nun in der Verteilung der Veränderung niederschlagen? Die Breite der Verteilung würde gleich bleiben, allerdings wäre sie auf der Abszisse vom Null-Punkt weg verschoben (in Richtung ‚krank' oder ‚gesund').

Welche Auswirkungen haben diese verschiedenen Verteilungsmuster nun auf die Korrelationen zwischen den verschiedenen Zeitpunkten? Die rechte Teilabbildung der Abb. 6.3 zeigt dies bei verschieden breiten Verteilungsmustern. Die oberste Linie der

[1] Dies Modell von Veränderung orientiert sich an hierarchischen Modelle, vgl. Gelman et al. (1997), Sullivan et al. (1999), Pinheiro & Bates (2000).

Teilabbildung zeigt den Verlauf an Korrelationen bei der engsten der möglichen Verteilungen von Veränderungen: In diesem Sample sind keine Personen vorhanden, die sich systematisch ändern. Diese Linie ist also identisch mit dem Verlauf von Korrelationen wie er in der rechten Teilabbildung der Abb. 6.2 auf Seite 53 wiedergegeben ist. Hier zeigt sich keine Niveauänderung der Korrelationen, sondern nur zufällige Schwankungen. Lässt man jedoch sukzessive mehr Personen im Sample zu, die sich wirklich ändern, werden die Kurven pro Zeitpunkt immer flacher. Die untersten Kurven zeigen auch, was passieren würde, wenn man den Zeitstrahl weiter ausdehnen würde: Mit Ausnahme der obersten Kurve (die würde bis ins Unendliche ihr Niveau behalten), nähern sich alle Kurven asymptotisch dem Null-Punkt (Korrelation von Null) an. Anzumerken ist weiterhin, dass auch diese Kurven, neben ihrem allgemeinen Trend asymptotisch zum Null-Punkt zu streben, zufallsbedingte Schwankungen beinhalten. Wie schon erwähnt, dürfen diese Verläufe nicht mit Reliabilitäten verwechselt werden, diese beinhalten per Definition keine wirklichen Änderungen, auch wenn dies vielleicht nur in einer denkbaren idealen Welt zutreffen mag.

Da die Korrelation zwischen verschiedenen Zeitpunkten ein Hauptbestandteil der Formel zum reliable change ist, lässt sich aus dem oben beschriebenen Effekt ohne Änderung der klassischen Formel (siehe Formel 2.2 auf Seite 21) eine Korrektur der ‚normalerweise zu erwartenden Veränderung' einarbeiten.

6.3 Fragestellung und Hypothesen

Die Fragestellung der vorliegenden Arbeit ist folglich, ob sich die oben durchgeführten Überlegungen zu den Implikationen und Auswirkungen der Anwendung einer solchen Korrektur bewahrheiten. innerhalb einer klinischen Untersuchung. Es steht also weniger eine mit ja oder nein zu beantwortende Forschungsfrage im Mittelpunkt, als das Aufzeigen von bestimmten Strukturen. Diese Strukturen beinhalten etwa die angenommenen Aussagen zur Veränderung (siehe oben) und die angenommenen Auswirkungen auf die Korrelationen zwischen den Zeitpunkten. In Frage steht also auch ein bestimmtes Veränderungsmodell, bzw. die Frage, was sich durch Anwendung dieses ‚zeigt'. **Hauptziel** ist die Einschätzung, welche Auswirkungen der modifizierte RC-Ansatz in der klinischen Forschung haben kann. Notwendiges Handwerkszeug zur Bewältigung dieser Forschungsfrage ist eine Datenlage aus der klinischen Psychologie. Es wird, um notwendigen Kenngrößen zur ‚normalerweise zu erwartenden Veränderung' zu erhalten, eine Normalstichprobe, sowie um die Anwendung des Modells in der klinischen Forschung zu demonstrieren, eine klinische Stichprobe benötigt.

Die Hypothesen sind in drei thematische Hypothesenkomplexe gegliedert. Der erste Komplex beinhaltet Hypothesen zu dem zugrundeliegenden Veränderungsmodell, der zweite Komplex zu dem Verlauf von Zusammenhängen (Korrelationen) abhängig vom zeitlichen Abstand, der dritte Komplex zu den neu zu bildenden RC-Indizes und der letzte, vierte Komplex zum Zusammenhang der neuen RC-Indizes mit alternativen Erfolgsmaßen. Der letzte Schritt kann als Validierungsschritt der RC-Indizes mit äußeren Kriterien genommen werden. Konkret ergeben sich also folgende Hypothesen:

Hypothesenkomplex 1: Veränderungsmuster. Dieser Hypothesenkomplex gliedert sich auf in Hypothesen zur Struktur der Änderungen und zu den resultierenden Korrelationsmustern.

Hypothesenkomplex 1.1: Struktur der Änderungungen. Analog zu den obigen Überlegungen werden verschiedene Veränderungsmuster erwartet:

1. In der Normalstichprobe wird im Durchschnitt über die verschiedenen Zeitpunkte hinweg keine Änderung erwartet.
2. In der behandelten, klinischen Stichprobe wird im Durchschnitt im Behandlungszeitraum eine in Richtung ‚Gesundung' verschobene Verteilung der Veränderungen erwartet.
3. Hinsichtlich verschiedener Merkmale werden sowohl bei der klinischen als auch bei der Normalstichprobe unterschiedlich breite Verteilungen von Veränderungen erwartet.

Hypothesenkomplex 1.2: Zeitlicher Verlauf der Korrelationen. Es werden spezifische strukturelle Eigenschaften der Korrelationen über die Zeit erwartet:

1. Der Ausgangspunkt der Korrelationskurven, d.h. der Punkt, an dem der geringste zeitliche Abstand besteht, wird sehr ähnlich dem zeitunabhängigen Schätzer der Reliabilität, Cohen's α sein.
2. Es wird erwartet, dass sich die Korrelationen mit zunehmendem zeitlichen Abstand verringern. D.h., dass die Annahmen der klassischen Testtheorie für längere Zeiträume zwischen den Messungen nicht zu halten sind.
3. Diese Verringerung wird asymptotisch die Null-Korrelation erreichen.
4. Über verschiedene Merkmalsbereiche hinweg wird die Abnahme der Korrelationen verschieden stark ausgeprägt sein.
5. Auch auf Itemebene werden sich unterschiedliche Korrelationskurven über die Zeit finden.
6. Die Ausgangspunkte der Korrelations-Kurven werden sich hinsichtlich verschiedener Merkmalsbereiche und über die Items hinweg unterscheiden (Retestkorrelation).

Hypothesenkomplex 2: Der modifizierte RC-Ansatz. In diesem Hypothesenkomplex geht es um die beschriebenen, neuen Kennwerte und die Auswirkungen bei der Anwendung dieser. Bei diesen wird in den ursprünglichen RC-Formeln statt der einfachen Retestkorrelation jeweils die der verstrichenen Zeitspanne entsprechende Retestkorrelation eingefügt.

1. Im Gegensatz zu dem klassischen Ansatz wird eine Verringerung der Erfolgsrate bei dem modifizierten Ansatz, d.h. mit der zeitlichen Komponente, erwartet.

2. Dies wird nicht erwartet, wenn der reliable-weighted Ansatz (siehe auch 2.3.3 auf Seite 21) in die Berechnung des modifizierten Ansatzes eingeht.
3. Die obigen beiden Hypothesen werden auch für die diagnostischen Untergruppen gelten.

Hypothesenkomplex 3: Validierung der RC-Indizes Bei diesem Hypothesenkomplex werden die gebildeten RC-Indizes mit alternativ formulierten Erfolgsmaßen verglichen. Hier wird insbesondere auf eine Substichprobe der Patienten zurückgegriffen, bei der eine Gewichtszunahme im Vordergrund steht. Es werden folgende dynamische und statische Erfolgsmaße (EM) definiert[1]:

dynamische EM: Zunahme des Body-Mass-Index pro Zeiteinheit. Dabei wird die Zeit sowohl als absolute (Behandlungsdauer) als auch als , therapeutische' (Therapiemenge) genommen.

statische EM: Hier werden die offiziellen Erfolgsmaße der Multizentrischen Studie, welche auf die DSM-III-R Diagnose bezogen sind, verwandt. Diese sind auf die Diagnose der Essstörung bezogen. So wird von einem Erfolg gesprochen, wenn alle Kriterien der jeweiligen Essstörung nicht mehr vorliegen, von einer Besserung wenn ein Kriterium weggefallen ist und von einem Mißerfolg wenn alle Kriterien weiterhin bestehen.

Folgende konkrete Hypothesen lassen sich aus den obigen Ausführungen generieren:

- Generell wird zwischen dem EDI und dem BMI bei Aufnahme und Entlassung ein moderater Zusammenhang angenommen.
- Es wird generell ein moderater Zusammenhang zwischen den durch den EDI definierten RC-Indizes und den alternativen Erfolgsmaßen angenommen.
- Es wird angenommen, dass die auf Zeit basierende RC-Maße mit anderen, dynamischen Erfolgsmaßen einen größeren Zusammenhang haben als mit statischen.

[1] Eine detaillierte Beschreibung der Maße findet sich im Ergebnisteil der Arbeit (siehe auch Abschnitt 8.3 auf Seite 71).

7 Methode

7.1 Stichproben

In der vorliegenden Arbeit wurden zwei getrennt voneinander erhobene Stichproben verwendet. Die erste Stichprobe, im folgenden *Normalstichprobe* genannt, wurde im Raum Stuttgart erhoben. Bei der zweiten Stichprobe handelt es sich um die Daten der bundesweit erhobenen multizentrischen Essstörungsstudie MZ-ESS (Kächele, Kordy, & Richard, 2001; Kächele, 2000, 1999). Die Normalstichprobe wurde vom Autor maßgeblich erhoben. Die Essstörungsstichprobe wurde von einer Arbeitgruppe, bei der der Autor zu der damaligen Zeit Mitglied war, erhoben. Die Nutzung der Daten wurde durch das geschäftsführende Gremium durch einen Antrag genehmigt.

7.1.1 Normalstichprobe

Die Normalstichprobe wurde im Raum Stuttgart an initial 300 Frauen aus 12 verschiedenen Einrichtungen (Schulen, verschiedenen Firmen, Universität) durch Vorlegen des EDI (Eating Disorder Intventory) in deutscher Übersetzung (Thiel & Paul, 1988) gewonnen. Erhoben wurde die Stichprobe im Zeitraum vom 13. Januar 1996 bis zum 8. Januar 1998. Die Daten des ersten Messzeitpunktes wurden zwischen dem 13. Januar 1996 und dem 27. Januar 1997 erhoben. Die meisten untersuchten Frauen wurden an der Universität rekrutiert (N=48). 38 Frauen waren Hebammenschülerinnen, 32 Frauen waren bei der Telekom beschäftigt, 15 waren bei der Kassenärztlichen Vereinigung Stuttgart beschäftigt und 11 Frauen besuchten zu dem Zeitpunkt die Schule. Die weiteren 144 Frauen wurden über verschiedene andere Wege für die Studie gewonnen (z.B. durch Mundpropaganda, bei Großveranstaltungen angesprochen, etc.). Tabelle 7.1.1 gibt einen Überblick über die Messzeitpunkte, die geplante Stichprobengröße pro Messzeitpunkt sowie das erreichte N.

Messzeitpunkt:	1	2	3	4	5	6	7
N-geplant	300	300	150	150	300	300	300
Wochen	0	2	4	8	16	26	52
Erreicht und Auswertbar	295	243	111	104	91	81	130

Tabelle 7.1: Stichprobengröße in der Normalstichprobe.

Die Stichprobengröße nahm mit weiteren Messzeitpunkten kontinuierlich ab (in Abbildung 7.1 auf der nächsten Seite ist das N abgetragen). Weiterhin wurden Bögen oft

erst nach der dritten Erinnerung zurückgesandt. Die Abbildung 7.1 zeigt die Verteilung der Messzeitpunkte als Boxplots. Auf der Abszisse der Abbildung sind der Messzeitpunkt abgetragen, die Ordinate zeigt die aus dem Rücksendeverhalten resultierenden Zeiträume. In dieser Darstellung ist neben der üblichen Boxplot-Statistik (Interqartile-

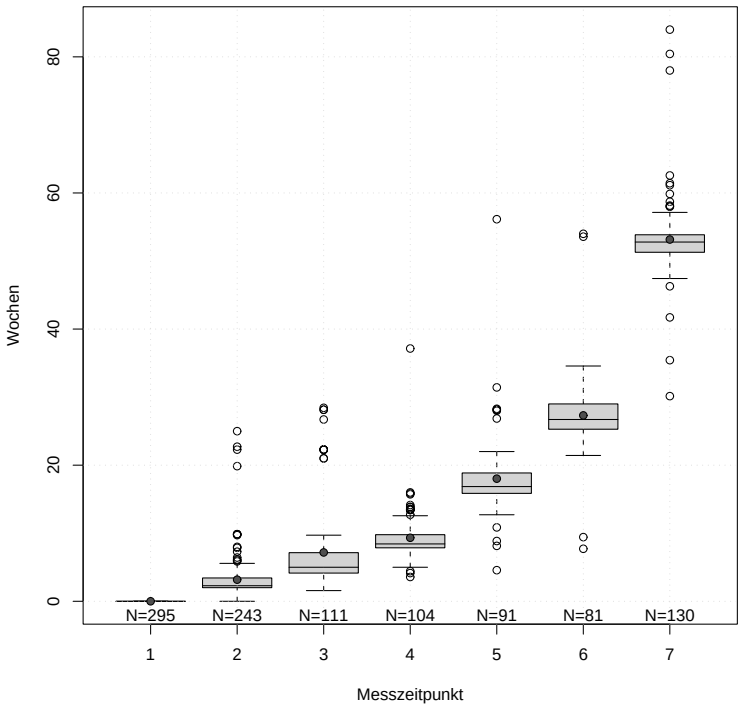

Abbildung 7.1: Verteilung der eingegangenen Bögen auf der Zeitachse

Range, Median, vgl. auch Cleveland, 1993) auch der Mittelwert als Punkt abgetragen. Da die Anfangspunkte der Untersuchung für verschiedene Personen unterschiedlich sind, wurde für jede Person ein individueller Nullpunkt auf der Zeitachse definiert (vgl. auch Abb. A.1 auf Seite 102 im Anhang). Wie in Abbildung 7.1 ersichtlich, überlappen sich die Verteilungen der Messzeitpunkte, d. h. für ein Großteil der Personen konnte das Erhebungsschema nicht realisiert werden.

7.1.2 Klinische Stichprobe

Die *klinische Stichprobe* wurde im Zeitraum von September 1993 bis Oktober 1995 in 43 tiefenpsychologisch orientierten Kliniken[1] in Deutschland erhoben. Weiterhin wur-

[1]Teil der Absprache zwischen den Kliniken war, dass keine klinikspezifischen Angaben zu den Patienten gemacht werden sollten. Im folgenden wird auch deshalb ausschließlich auf die Gesamtstichprobe

de bis Ende 1998 die 2.5-Jahres Katamnese erhoben. Einschlußkriterium war eine Essstörungsdiagnose nach DSM-III-R (Anorexie, Bulimie und Doppeldiagnose) und die Forderung, dass die Patientinnen während des stationären Aufenthaltes das 18. Lebensjahr bereits vollendet hatten. Abbildung 7.2 gibt den Aufbau der Studie schematisch wieder. In der vorliegenden Arbeit werden aufgrund der Fragestellung jedoch nur die Datenpunkte ‚Aufnahme' und ‚Entlassung' verwandt.

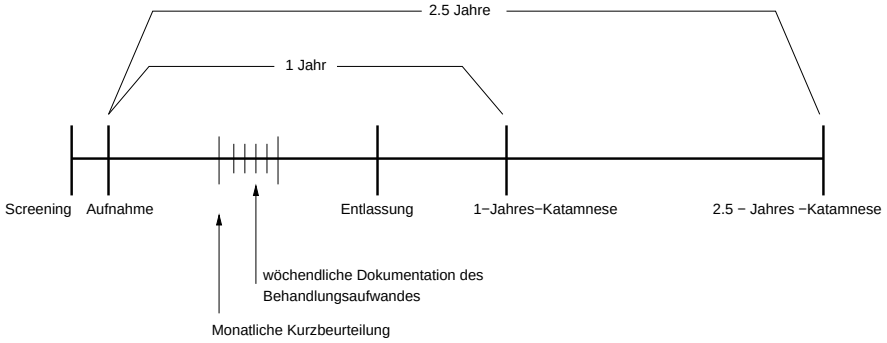

Abbildung 7.2: Design der klinischen Stichprobe

Von den in allen Kliniken initial gescreenten Patientinnen wurde von 1247 Patientinnen die Behandlungsphase dokumentiert. Eine detaillierte Dokumentation der insgesamt gescreenten und der dabei von der Teilnahme an der Studie ausgeschlossenen Patientinnen sowie der Gründe des Ausschlusses wurde nur in einigen wenigen Kliniken durchgeführt. Der Anteil der Patientinnen, der von den einzelnen Kliniken in die Studie eingebracht wurde, ist unterschiedlich und reicht von einem Fall bis zu 246 Fällen. Von den 1247 Patientinnen wurden 76 (6.1 %) aus der weiteren Datenverarbeitung sowie den beiden katamnestischen Erhebungen ausgeschlossen. Das Fehlen der Diagnosestellung zu Behandlungsbeginn war mit n=31 die häufigste Ursache des Ausschlusses. Protokollverletzungen, die hauptsächlich das bei kurzen Behandlungen gar nicht oder nicht termingerechte Ausfüllen der Kurzbeurteilungen bedeuteten, waren mit n=14 ein weiteres Ausschlusskriterium. Fehlende Gewichtsangabe bei Anorexie (n=8), widersprüchliche Bulimie-Diagnose (n=4), zu frühe Abgabe des Therapeuten- und/oder Patientenentlassungsbogens (n=17) sowie das Nichterfüllen des Alterskriteriums bei Eintritt in die Studie (n=2) waren weitere Ausschlussgründe. Somit verblieben 1171 Patientinnen zur weiteren Datenerhebung und -verarbeitung. Davon waren 29 (2.5 %) männliche und 1142 (97.5 %) weibliche Patientinnen. Um die Studie bei der Vielzahl der aus der MZ-ESS resultierenden und veröffentlichten wissenschaftlichen Arbeiten einheitlich darzustellen, wurde mit allen an der Studie teilnehmenden Institutionen und deren Mitarbeitern vereinbart, die Stichprobe auf Basis der Therapeutendiagnose vorzustellen und zu beschreiben. Nach Beurteilung der Therapeuten erfüllten 355 (30.3 %) Patientinnen die Kriterien der Anorexie, 647 (55.3 %) die der Bulimie und

eingegangen.

169 (14.4 %) die einer Doppeldiagnose. Zu Beginn der Indexbehandlung waren die Patientinnen zwischen 17.5 und 56.8, im Durchschnitt 25.5 Jahre (sd= 6 Jahre) alt. Insgesamt nahmen mehr jüngere Patientinnen an der Studie teil; nur 25 % waren über 28.3 Jahre alt. Diese Verteilung ist in allen Diagnosegruppen vergleichbar. Im Durchschnitt wurde mit 18.7 Jahren (sd= 5.0) eine Essstörung das erste Mal diagnostiziert, so dass die Patientinnen im Mittel bereits 7.3 (± 6) Jahre essgestört waren, bevor sie in die Studie aufgenommen wurden.

Die Behandlungszeit war insgesamt weniger von Patientenmerkmalen als von den behandelnden Kliniken abhängig (Kächele et al., 2001). Abbildung 7.3 zeigt die große Variabilität der Behandlungsdauern.

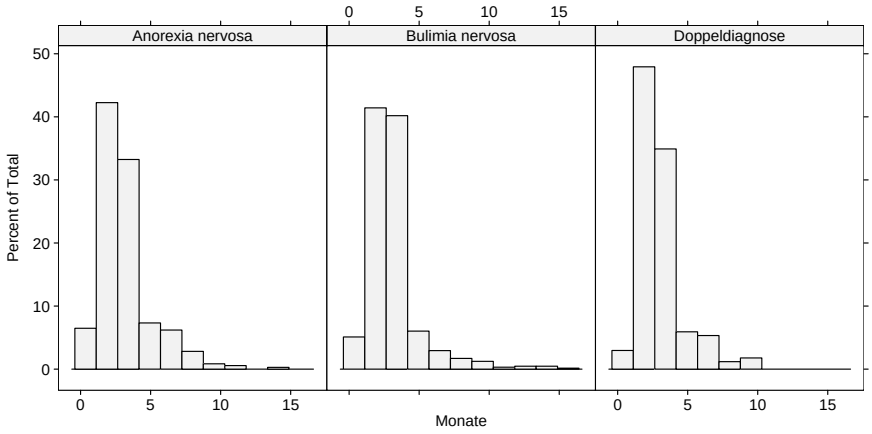

Abbildung 7.3: Verteilung der Aufenthaltsdauer abhängig von der gestellten Diagnose.

Wie aus den Abbildungen ersichtlich, sind diese nicht abhängig von der Diagnose nach DSM-III-R[1].

Tabelle 7.1.2 auf der nächsten Seite gibt einen Überblick zur Verteilung des BMI über die Diagnosegruppen. Wie erwartet, unterscheidet sich die Gruppen hier (Anova: $F = 95.48$, $DF = 1$ & 1168, $p <= 0.0001$).

7.2 Instrumente

Das in dieser Arbeit hauptsächlich verwendete Instument ist der *Eating Disorder Inventory* (Abkürzung: EDI). Das englische Originalinstrument wurde von Garner et al. (1983, 1985), Garner & Olmsted (1984), Garner (1991) entwickelt. Aus einem durch Expertenbeurteilungen gebildeten Itempool wurden diejenigen Items ausgewählt, welche gut zwischen Patientinnen mit Anorexia nervosa und einer Kontrollgruppe differenzierten. Die zugrundeliegenden Skalen wurden a priori konstruiert.

[1]Dies lässt sich auch via multivariate Varianzanalyse zeigen: Weder Aufenthaltsdauer noch Katamnesedauer sind abhängig von der Diagnose: ($Phillai = 0.003, appr.\ F. = 0.3, Pr(> F) = 0.99$).

Statistik	Anorexia nervosa	Bulimia nervosa	Doppeldiagnose
Minimum	9.82	13.54	11.90
1. Quantile	13.87	19.31	15.57
Median	15.03	21.48	16.86
Mittelwert	15.11	22.70	17.16
Standardabweichung	1.92	5.52	3.39
2. Quantile	16.41	24.05	18.07
Maximum	21.30	51.68	44.22

Tabelle 7.2: Überblick über den BMI zum Aufnahmezeitpunkt bei den drei Diagnosegruppen.

Mittlerweile liegt der EDI in zwei deutschen Übersetzungen vor. In der zweiten Version (Thiel et al., 1997) wurden einige Items neu formuliert und die ursprünglich 64 Items wurden um 27 Items ergänzt. In der vorliegenden Arbeit wurde jedoch der sogenannte ‚EDI-Münster' (Meermann & Vandereycken, 1987) verwandt, insbesondere auch deshalb, da in der Planungsphase der Studien die revidierte Fassung (Thiel & Paul, 1988) noch nicht verfügbar war. Der EDI besteht aus 64 Selbstschilderungen von Verhaltensweisen oder Kognitionen, die jeweils auf sechs Stufen von ‚nie' (1) bis ‚immer' (6) zu beantworten sind. Sie verteilen sich auf die acht a-priori-Skalen. Eine vollständige Liste der Items ist im Anhang im Abschnitt A.2 auf Seite 102 wiedergegeben.

drive for thinness (7 Items): Beispielitem: ‚Ich esse Süssigkeiten und Kohlehydrate, ohne nervös zu werden'. Dies ist die eigentliche Anorexie-Skala. Sie drückt vor allem die Furcht aus, Gewicht zuzunehmen oder, anders gewendet, den Drang dünn zu sein. Cronbachs Alpha betrug in der von den Autoren beschriebenen klinischen Stichprobe (KSP) von 205 Frauen mit Bulimia nervosa 0.77, in den Vergleichsstichproben (VST) nicht erkrankter Personen (183 Frauen, 104 Männer) war Cronbachs-Alpha 0.82 für die Frauen und 0.73 für die Männer (Thiel & Paul, 1988).

bulimia (7 Items): Beispielitem: ‚Ich stopfe mich mit Essen voll'. Diese Skala drückt vor allem die mit dem Syndrom der Bulimie verknüpften Kognitionen aus (Cronbachs Alpha: klinische Stichprobe (KST): 0.72, Vergleichsstichprobe (VST): 0.72 und 0.59).

body dissatisfaction (9 Items): Beispielitem: ‚Ich denke, meine Oberschenkel sind zu dick'. Diese Sklala faßt insbesondere körperbezogene Kognitionen zusammen. (Cronbachs alpha: KSP: $0,89$ VST: $0,89$; $0,79$).

ineffectiveness (10 Items): Beispielitem: ‚Ich wünschte, ich wäre jemand anders'. Diese Skala hat eine starke konzeptuelle Ähnlichkeit mit der Selbstwirksamkeit und zu dem Konzept der Selbstwertregulation (Cronbachs alpha: KSP: $0,90$ VST: $0,81$; $0,80$).

perfectionism (6 Items): Beispielitem: ‚Ich hasse es, nicht der/die Beste zu sein'. Diese Skala drückt insbesondere überhöhte Ansprüche an sich aus (Cronbachs alpha: KSP: $0,77$ VST: $0,67$; $0,75$).

interpersonal distrust (7 Items): Beispielitem: ‚Es fällt mir schwer, meine Gefühle anderen gegenüber auszudrücken'. Diese Skala drückt das Misstrauen gegenüber anderen, die Gehemmtheit oder sozialen Rückzug aus. (Cronbachs alpha: KSP: $0,82$ VST: $0,75$; $0,69$).

introceptive awareness (10 Items): Beispielitem: ‚Ich fühle mich aufgebläht, wenn ich nur eine Kleinigkeit gegessen habe'. Diese Skala beschreibt insbesondere die körperbezogene Aufmerksamkeit (Cronbachs alpha: KSP: $0,84$ VST: $0,74$; $0,69$).

maturity fears (8 Items): Beispielitem: ‚Die besten Jahre des Lebens sind die, in denen man erwachsen wird'. Diese Skala beschreibt die Furcht, erwachsen zu werden (Cronbachs alpha: KSP: $0,79$ VST: $0,65$; $0,55$).

Eine durchgeführte Faktorenanalyse konnte die faktorielle Validität bestätigen. Durchgeführte Diskriminanzanalysen und T-Tests mit klinischen Stichproben konnten die Diskriminationsfähigkeit des Instrumentes bestätigen. Eine Normierung wurde für die deutsche Fassung nicht durchgeführt[1]. Angegeben werden jedoch: - Mittelwerte und Standardabweichungen aller Skalen für eine Bulimiestichprobe (N = 205 Frauen) und der Vergleichsstichprobe. Die Skalenwerte werden als Mittelwerte der zugrundeliegenden Items berechnet, so dass der Wertebereich der Skalen denen der Items entspricht. Hierbei ist allerdings zu beachten, dass die unaufälligen Antwortabstufungen 1-3 zum Wert 0 zusammengefaßt werden und die Antwortmöglichkeit (6) den Wert 3 erhält, die Antwortmöglichkeit (5) eine 2 usw. Der Vorteil dieser Auswertungsstrategie liegt in der klinischen Bedeutsamkeit, nicht auffällige Werte sind generell Null. Bzgl. der Forschung ist diese Skalenberechnung jedoch eher von Nachteil, da praktisch die Verteilung der Antwortmöglichkeiten abgeschnitten wird.

Weiterhin wurde der Body-Mass-Index verwandt. Es handelt sich dabei um ein hinsichtlich der Körpergröße korrigierte Gewichtsmessung. Dieser berechnet sich wie folgt: $BMI = \frac{kg}{m^2}$.

7.3 Statistische Analyse

Die gesamte statistische Analyse wurde mit dem Programmpacket R (Ihaka & Gentleman, 1996) durchgeführt. Die graphische Auswertung orientiert sich stark an Cleveland (1994, 1993). Besondere Berücksichtigung fand die NLME - Bibliothek von Pinheiro & Bates (2000), welche u. a. besonders die Modellierung individueller Verlaufskurven, in den sogenannten ‚random effects models', geeignet ist.

[1] Kordy, Percevic, & Martinovich, 2001 haben zwar dargelegt, dass es sinnvoll ist länderspezifische Normen zu etablieren, jedeoch werden diese in dieser Arbeit nicht genannt.

Grundlage, im Sinne des sine qua none, für die weitergehende Analyse war die Modellierung der Korrelationen zwischen den verschiedenen Zeitpunkten. Eine der Hauptannahmen lautet ja, dass sich hier ein bestimmtes Muster finden sollte (wenn die zugrundeliegenden Modellannahmen richtig sind). Das heißt, der Verlauf der Korrelationen über die Zeit sollte nicht einfach zufälligen Schwankungen unterworfen sein. Hier ergibt sich somit unmittelbar die Frage, wie diese Modellierung durchzuführen ist.

Ein in der Vergangenheit eingeschlagener Weg (Malewski & Dillmann, 1997) bestand in der Darstellung destinkter Korrelationen zwischen den Zeitpunkten, also etwa zwischen Messzeitpunkt 1 und 2, Messzeitpunkt 1 und 3, etc. Wie jedoch schon aus der Abbildung 7.1 ersichtlich, findet sich in der Normalstichprobe eine starke Variation der Fallzahlen über die Messzeitpunkte, welche im starken Maße vom ursprünglichen Erhebungsplan (vgl. auch Tabelle 7.1.1 auf Seite 58) abweicht. Die Auswertungsstratgie, welche auf eine Restriktion auf die Fälle, die innerhalb dieses Schematas liegen, hat vor allem den Nachteil des hohen Informationsverlustes und die Unsicherheit, durch Abschneiden der Verteilung einen Bias zu produzieren (Little & Rubin, 1987). Weiterhin impliziert dieser Weg eine unnötige, willkürliche Entscheidung: Wie ist der Toleranzbereich zu bilden? Ist ein Messwert, der 10 Tage hinter der eigentlichen ‚Fälligkeit' liegt, zu verwerfen, oder etwa erst nach 14 Tagen? Oder noch grundsätzlicher: mit welcher Begründung ist ein außerhalb dieses Toleranzbereichs liegender Wert ein *Fehler*?

Im folgenden wurde ein alternativer Weg eingeschlagen, der die obigen Probleme löst. Analog zur Berechnung des gleitenden Durchschnitts bei Zeitreihendaten wird auch hier ein Fenster entlang der Zeitachse geschoben. In diesem Fenster liegen jeweils eine bestimmte Anzahl von Personen mit Messwerten zu je zwei Zeitpunkten, d. h. innerhalb des Fensters ist es möglich, eine Korrelation zu berechnen. Dabei wurde die Fenstergröße nicht über die Zeitspanne definiert, sondern über die Anzahl der Personen, die in diesem Fenster liegen sollten.

$$r_{t_{diff}} = r_{x_{t-T}, x_{t+T}} \tag{7.1}$$

D. h. wie in 7.1 dargestellt, ergibt sich die Korrelation für den Zeitpunkt t als die Korrelation der in der Zeitspanne $t - i \ldots t + i$ liegenden Personen.

8 Ergebnisse

8.1 Analyse der Veränderungsmuster

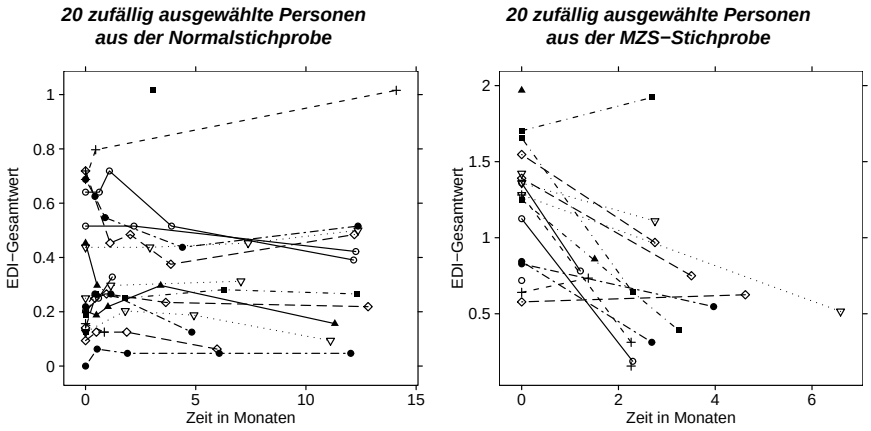

Abbildung 8.1: Beispiel verschiedener Verläufe in den beiden Stichproben

Beide Stichproben wurden zu mehreren Zeitpunkten untersucht. In dem folgenden Abschnitt werden die Differenzen der Personen über die Zeitpunkte hinweg analysiert (*Hypothesenkomplex 1.1* siehe Abschnitt 6.3 auf Seite 55). Ziel ist dabei eine Beschreibung dieser Veränderungsmuster. Abbildung 8.1 zeigt je 20 zufällig ausgewählte Personen aus den beiden Stichproben. Diese Grafik verdeutlicht noch einmal die Eigenheiten der beiden Stichproben und führt gleichzeitig zu dem folgenden, weiteren Analyseschritten. Betrachtet man diese Grafik, zeigt sich:

- Während auf der linken Seite, in der Teilabbildung der Normalstichprobe, die Verläufe durch mehrere Messzeitpunkte ‚gestützt' werden, sind auf der rechten Seite nur zwei Messzeitpunkte vorhanden[1], folglich ‚fehlt' in der klinischen Stichprobe die Information, was sich zwischen Aufnahme und Entlassung getan hat[2]

[1] In der gesamten Arbeit wird auf die Analyse der Katamnese verzichtet (zum Design der Studie vgl. Abb. 7.2 auf Seite 60).

[2] Jedenfalls bzgl. des eingesetzten EDI. Dieser wurde nur zu den Zeitpunkten Aufnahme, Entlassung sowie zur Katamnese verwandt.

- In beiden Stichproben findet sich eine Variabilität der zeitlichen Differenz zwischen erster und letzter Messung. So ist die Standardabweichung dieser Zeitspanne in der Normalstichprobe 0.37 Monate, der klinischen Stichprobe 1.8 Monate. Das bedeutet, dass die Messintervalle in der klinischen Stichprobe unterschiedlicher sind; in der Normalstichprobe dagegen homogener. Ein Test der Hypothese, dass beide Varianzen gleich sind, kann abgelehnt werden: das beobachtete Verhältnis der beiden Varianzen von 0.04 mit dem 95% Vertrauensbereichs von 0.03 bis 0.06 (p<=0.0001)[1] ist deutlich unterschiedlich von dem unter der Null-Hypothese erwarteten Varianzverhältnis von 1. Auch ist der zugrundeliegende Mechanismus, der diese Variabilität erzeugt, ein anderer: In der klinischen Stichprobe wird diese insbesondere durch die sehr unterschiedlichen Behandlungszeiten in den Kliniken bedingt, während es sich bei der Normalstichprobe vor allem durch die unregelmäßigen Rücksendedaten ergibt.

- Betrachtet man den Verlauf der durch Linien verbundenen Punkte, so fällt auf, dass die Mehrzahl der Verläufe in der klinischen Stichprobe eher fallen, während in der Normalstichprobe eine größere Heterogenität der Verläufe sichtbar wird, d.h. hier gibt es auf den ersten Blick ähnlich viele steigende wie fallende Linien.

Natürlich handelt es sich bei den obigen Punkten nur um erste Beobachtungen weniger Fälle. Analog zu diesen illustrierenden Beobachtungen sollen die Veränderungseigenschaften für die gesamten Stichproben aufgezeigt werden.

Zu diesem Schritt wird, da die Darstellung aller Verläufe, wie in der obigen Abbildung, zu unübersichtlich wäre, eine Abstraktion notwendig. Das heißt der Verlauf jeder einzelnen Person muss durch einige Maßzahlen repräsentiert werden. Diese Repräsentation, dass heißt die Abstraktion von den realen Datenpunkten, kann etwa dadurch geschehen, dass jeder Verlauf als eine Linie repräsentiert wird (lineare Regression).

Abbildung 8.2 auf der nächsten Seite gibt hierzu ein Beispiel. Die beobachteten Messwerte sind hierbei durch durchgehende Linien dargestellt, die lineare Vorhersage durch eine gestrichelte Linie. Es wird also von den beobachteten Werten abstrahiert und pro Person ein (linearer) Trend geschätzt. Diese Analyse wäre natürlich auch über alle Personen möglich. Ergebnis wäre dann die Änderung, hinsichtlich einer bestimmten Merkmales der Stichprobe in der Zeit, also etwa einer ‚Gesundung'. Im Kontext der random-effect-models wird eher dieser zweistufige Weg eingeschlagen. In den folgenden Ausführungen werden die Verteilungen der Kenngrößen dieser Trends dargestellt. Jede Regressiongerade kann durch zwei Kenngrößen eindeutig beschrieben werden: Der Punkt, in dem diese die Ordinate schneidet, also dem Intercept und durch den Punkt, der die Steigung der Linie angibt. Diese beiden Kenngrößen werden im Folgenden als geschätzter Ausgangspunkt und geschätzte Geschwindigkeit der Änderung benannt. So ist in der klinischen Stichprobe etwa der Ausgangspunkt die geschätzte Merkmalsausprägung bei Klinikaufnahme; die Geschwindigkeit der Änderung die Differenz der Merkmalsausprägung pro Zeiteinheit. In dieser Analyse wird erstmalig auch ein ‚Feh-

[1] F-Test (F = 0.0437, num df = 110, denom df = 1170, p-value = < 2.2e-16). Auch der nichtparametrische Ansari-Bradley Test (AB = 6956.5, p-value = < 2.2e-16) zeigt dies (Hollander & Wolfe, 1973).

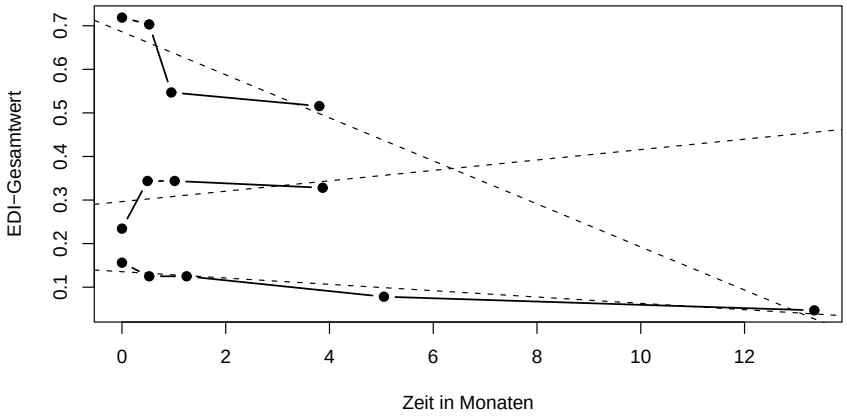

Abbildung 8.2: Beispiel für drei Verläufe deren jeweiligen linearen Vorhersagen.

ler‘ geschätzt. Dieser ist individuenspezifisch und, über alle Individuen, die Summe der Abweichungen der Messwerte der individuellen Personen von der jeweiligen geschätzten Kurve. D.h. es ergibt sich pro Person eine Abweichungen von der geschätzten Linie.

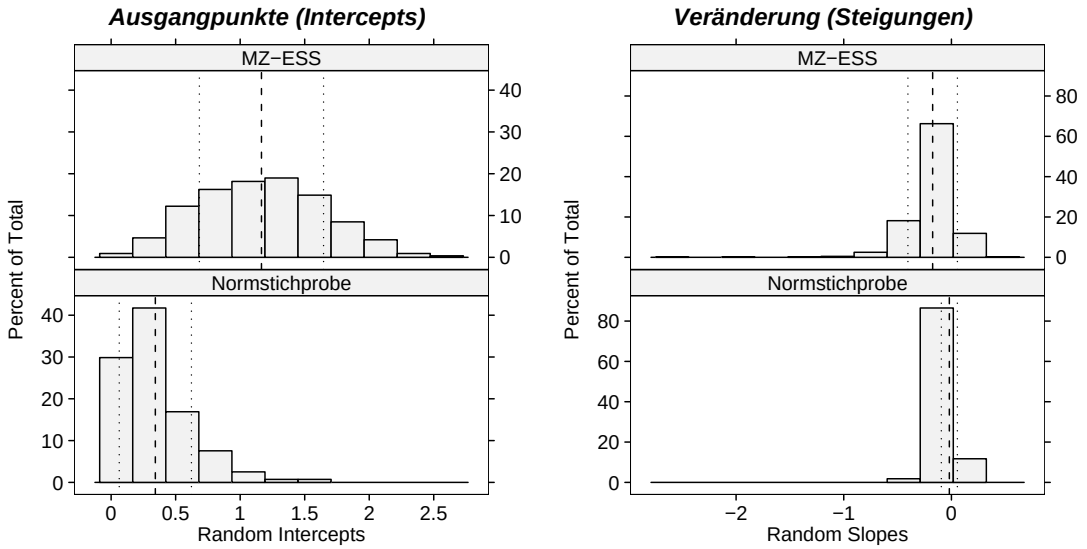

Abbildung 8.3: Vergleich der Veränderungswerte.

Die Abbildung 8.3 zeigt diesen Vergleich (MZ-ESS: Multizentrische Essstörungsstudie). Dargestellt wird die Verteilung durch sogenannte Histogramme, die in bestimm-

ten Intervallen, den Anteil der in diesen Intervallen liegenden Personen durch Balken darstellen. Auf der linken Seite ist die Verteilung aller geschätzten Intercepts, welche als die Ausgangspunkte der Personen angesehen werden können, abgetragen, auf der rechten Seite die Steigungen der Personen. Diese Steigungen können auch als Veränderungsgeschwindigkeit (Ausmaß der Veränderung pro Zeiteinheit) verstanden werden. Zusätzlich ist als Orientiertungspunkt der Mittelwert und die Standardabweichungen als vertikale Linien abgetragen.

Was zeigt diese Grafik? Wird die linke Seite der Abbildung 8.3 auf der vorherigen Seite betrachtet, wird deutlich, dass sich die Intercepts (Ausgangswerte) in den beiden Stichproben unterscheiden. Während diese im Durchschnitt in der klinischen Stichprobe bei 1.16 (SD=0.48, Min=0.06, Max=2.62) liegen, ist dies in der Normalstichprobe deutlich anders. Hier liegt der Durchschnitt der Intercepts bei 0.34 (SD=0.28, Min=0.02, Max=1.70). Das bedeutet, die klinische Stichprobe liegt deutlich in dem von der EDI-Gesamtskala definierten Bereich der Essstörung. Auch die Form der Verteilungen unterscheidet sich deutlich. Während die klinische Stichprobe symmetrisch ist (Skewness=0.17), ist die Verteilung der Normalstichprobe deutlich linksschief (Skewness=1.7).

Auch die Verteilungen der Steigungen der individuellen Verlaufskurven unterscheiden sich zwischen den Stichproben (rechte Teilabbildung). Während der Mittelwert der Normalstichprobe nahe Null liegt (Mean=-0.02), ist dieser in der klinischen Stichprobe von Null unterschieden (Mean= -0.17). Das bedeutet, besonders in der klinischen Stichprobe wird eine Änderung in die Richtung der ‚Gesundheit' deutlich. Auch die Variabilität der Steigungen in den beiden Stichproben unterscheidet sich. So findet sich in der Normalstichprobe mit einem Wertebereich von -0.60 bis 0.11 eine mehr als ein Drittel so kleine Standardabweichung (sd=0.07) wie in der klinischen Stichprobe (sd=0.23); und auch der Wertebereich der klinischen Stichprobe ist deutlich gespreizter (-2.62,0.51). In der klinischen Stichprobe sind 82.9% der Slopes kleiner Null, in der Normalstichprobe 67.57%.

Weiterhin findet sich ein recht deutlicher Zusammenhang zwischen den Intercepts und Slopes. So betrug die Korrelation zwischen den beiden Kennzahlen in der Normalstichprobe -0.18 (95% Pearson-Konfidenzintervall:-0.3, -0.05; p=006) und in der klinischen Stichprobe -0.33 (95% Pearson-Konfidenzintervall:-0.39,-0.27; p<=0.0001). Das bedeutet, dass je höher die initialen, geschätzten Ausgangswerte sind, desto niedriger sind auch die Slopes. Dies bedeutet, das die Änderungsrate abhängig vom Anfangswert ist.

8.2 Modellierung der von der Zeit abhängigen Korrelationen

Ziel dieses Abschnittes ist die Modellierung der sich mit fortschreitender (*Hypothesenkomplex 1.2* siehe Abschnitt 6.3 auf Seite 55) Zeit ändernden Korrelationen. Insbesondere wird angenommen, dass die Korrelationen zwischen dem Zeitpunkt t_1, t_2 kleiner einer Korrelation t_1, t_3 ist, d.h. mit Vergrößerung der zeitlichen Differenz sollte die Korrelation zwischen den Testwerten abnehmen. Das wäre auch so ausdrückbar: Gruppen von Personen sind sich im Vergleich zu gestern ähnlicher als wie von vor zwei Jahren. Das

eingesetzte Verfahren ist ausführlich unter dem Abschnitt 7.3 auf Seite 63 beschrieben.

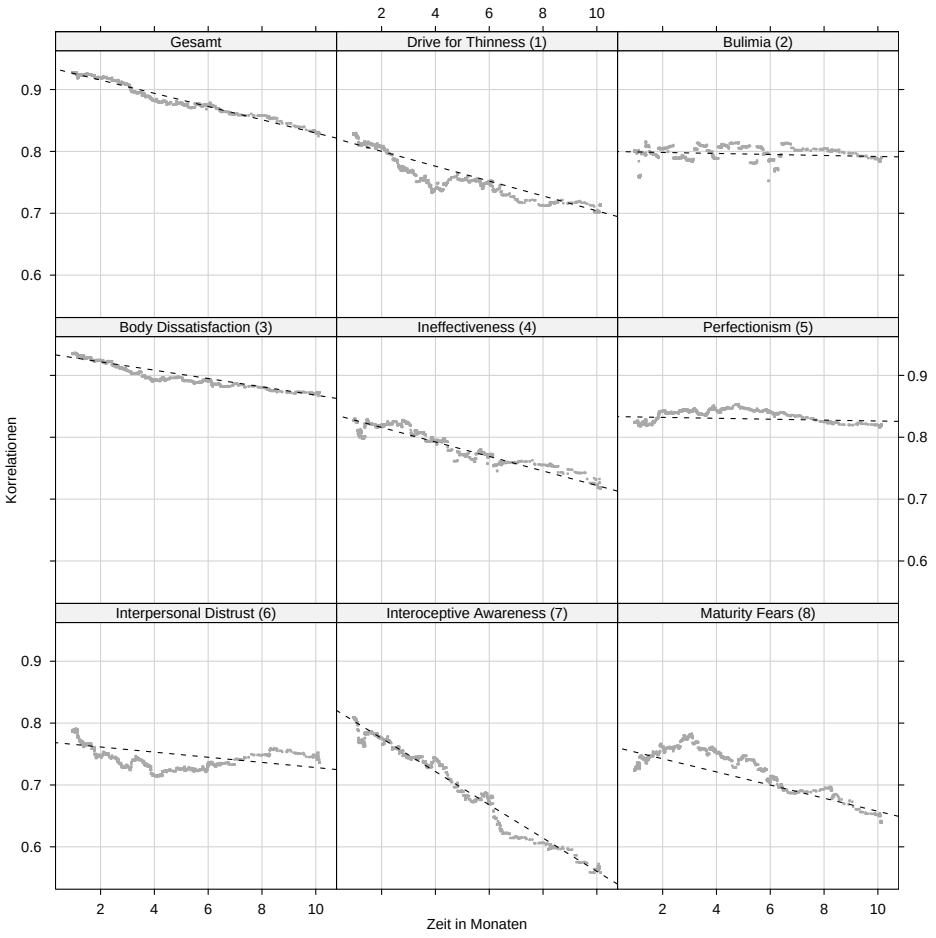

Abbildung 8.4: Test-Retestkorrelation mit zunehmendem zeitlichen Abstand zwischen den Messungen

Abbildung 8.4 zeigt die nach dem oben beschriebenen Verfahren gewonnen Ergebnisse für die verschiedenen Skalen des EDI. In diesen Grafiken sind die Korrelationen abhängig von der Zeitspanne, die zwischen den Messungen liegt, dargestellt. Zusätzlich sind die berechneten linearen Regressionen $Cor(x_t, x_{t+i}) \sim Diff(t, t+1)$ als gestrichelte Linien in die Grafik eingetragen. Analog zum vorherigen Abschnitt lassen sich auch hier die Verläufe zusammenfassen und die Verteilungen dieser zusammengefassten Verläufe analysieren. Der Unterschied zum vorherigen liegt jedoch darin, dass nicht individuelle Verlaufskurven betrachtet werden, sondern Verläufe von Korrelationen. Als erster Schritt ist jedoch auch hierbei die Abstraktion von den konkreten Kurven notwendig. Dieser geschieht durch die Schätzung eines linearen Trends, da sich durch die graphische Exploration keine Hinweise auf Nichtlinearität finden lassen. Eine (klei-

Kennwerte

	Alpha	Retest-Korrelation (1-2)	Intercept	Steigung
Gesamt	0.942	0.918	0.937	−0.011
Drive for Thinness	0.866	0.816	0.824	−0.012
Bulimia	0.860	0.829	0.800	−0.001
Body Dissatisfaction	0.945	0.933	0.935	−0.007
Ineffectiveness	0.901	0.878	0.839	−0.012
Perfectionism	0.781	0.814	0.834	−0.001
Interpersonal Distrust	0.825	0.809	0.770	−0.004
Interoceptive Awareness	0.858	0.833	0.829	−0.027
Maturity Fears	0.753	0.616	0.764	−0.011

Korrelationen der Kennwerte

	Alpha	Retest-Korrelation (1-2)	Intercept
Retest-Korrelation1.2	0.886		
Intercept	0.824	0.794	
Steigung	-0.166	0.006	-0.079

Tabelle 8.1: Retestreliabilität (Intercept), Veränderungspotential (Steigung) und Cronbachs Alpha in der Normalstichprobe

nere) Ausnahme stellt die Skala ‚Interpesonell Distrust' dar. Hier zeigt sich mit zunehmender Zeitspanne zwischen den Messungen, nach einem gewissen Abfall, ein weiterer Anstieg der Korrelationen. Zwischen den Skalen bestehen deutliche Unterschiede, sowohl bezüglich der Ausgangpunkte als auch bezüglich der Stärke des Abfalles. So nimmt die Korrelation bei den Skalen 2, 8 und 9 stark ab, während die anderen Einzelskalen wie die Gesamtskala einen eher moderaten Abfall der Korrelationen zeigen. Hinsichtlich der Ausgangspunkte findet sich ebenfalls eine beträchtliche Variabilität.

Tabelle 8.1 gibt die Kennzahlen der geschätzten linearen Regressionen wieder (die Spalten ‚Intercept' und ‚Steigung') sowie auch die berechneten Cronbachs-Alpha Koeffizienten zum ersten Messzeitpunkt und die Korrelation zwischen den ersten beiden Messzeitpunkten in der Normalstichprobe. Der Ausgangspunkt des Abfalls der Korrelation mit Vergrößerung der dazwischenliegenden Zeitspanne, der Intercept, variierte erheblich zwischen den Skalen. So liegt der Range dieser Kennzahlen zwischen 0.76 (Maturity Fears) und 0.93 (Gesamtskala), der Durchschnitt dieser Parameter bei 0.83 (sd=0.06).

Da der Ausgangspunkt der geschätzten Regressionen auch als die eigentliche Retestkorrelation interpretierbar ist, ist es hier insbesondere auch sinnvoll, Cronbachs Alpha und die Korrelationen zwischen den ersten beiden Messzeitpunkten mit darzustellen (Retestreliabilität). Hier zeigt sich, dass die Kennzahlen der geschätzten Intecepts bis auf die Skala Perfectionism kleiner als Cronbachs Alpha und die Retestreliabilitäten

sind. Die Differenzen zwischen Cronbachs Alpha und den Intecept variieren zwischen -0.05 und 0.06 (sd=0.04, mean=0.02), zwischen den Retestreliabilitäten und den Intercepts zwischen -0.15 und 0.04 (sd=0.06, mean=-0.01). Die Zusammenhänge zwischen den einzelnen Kennwerten sind recht hoch, am geringsten ist der Zusammenhang zwischen der Retestkorrelation und den Intercepts (0.79).

Ein Teil dieser Tabelle ist weiterhin zur besseren Veranschaulichung als Abbildung realisiert. So zeigt der obere Teil der Abbildung 8.5 auf der nächsten Seite die in der Tabelle 8.1 auf der vorherigen Seite als Intercept und Steigung gekennzeichneten Spalten. Die untere Teilabbildung zeigt hingegen die nach dem selben Verfahren geschätzten Kennzahlen der Items. Zugrundeliegend sind lineare Regressionen, bei denen die Korrelation zwischen den Zeitpunkten abhängige Variable, die Zeit die unabhängige Variable war. Die obere Subgrafik zeigt die Skalen, die untere die Items.

Abbildung 8.5 auf der nächsten Seite visualisiert in der oberen Teilabbildung somit die Lokalisisation der einzelnen Skalen. Auf der Abszisse sind hier nochmals die Intercepts abgetragen, auf der Ordinate die Slopes. Diese Teilabbildung soll jedoch nicht allein zur weiteren Veranschaulichung dieser Verhältnisse dienen, sondern soll vor allem einen Vergleich mit den geschätzten Lokalisationen der Items der Skalen in diesem imaginären Raum ermöglichen. Die untere Teilabbildung zeigt hierbei, analog zur Darstellung der Skalen, die geschätzten Kennwerte auf Itemebene. Es fällt hierbei auf, dass die Schwerpunkte der Item-Punkte-Wolken nur teilweise der Verteilung der Skalen auf diesem zwei dimensionalen Raum entsprechen[1].

8.3 Verschiedene Berechnungsweisen des Reliable-Change-Begriffs

Die beiden vorhergehenden Analyseschritte sind die Voraussetzung für diesen, die eigentlich zentrale Fragestellung (*Hypothesenkomplex 2* siehe Abschnitt 6.3 auf Seite 55) der Arbeit beleuchtenden Schritt. Hier geht es insbesondere darum, die Differenz zwischen einem, die zeitliche Variabilität berücksichtigen Ansatz des Reliable change Begriffs[2] und den klassischen Versionen des dieser Umsetzung aufzuzeigen. Der Reliable-Change-Begriff ist ein Urteil, dass bestimmt, ob die Veränderung einer Person a) eine reliable Verschlechterung darstellt, b) keine reliable Veränderung feststellbar ist und c) ob es sich um eine reliable Verbesserung des Zustandes handelt. D. h. hinsichtlich einer Gruppe von Personen ergeben sich Prozentangaben, z.B. ‚17% der Personen haben sich gebessert'. Alle Berechnungen der RC-Indizes beziehen sich auf ein 5% Signifikanzniveau.

Tabelle 8.3 auf Seite 74 gibt die prozentualen, reliablen Verbesserungen der klinischen Stichprobe wieder. Es werden hierbei verschiedene Berechnungsweisen verglichen:

[1] Eine derartige Analyse ermöglicht, was jedoch nicht Gegenstand dieser Arbeit ist, insbesondere die Unterscheidung veränderungssensitiver vs. trait-fokussierender Skalen. Mit einem derartigen Verfahren können Skalen zu extrahieren, ist immer eine Optimierungsaufgabe, die die sich zum Teil ausschließenden Aufgaben enthält, reliabelsten Items (Itercepts) oder die veränderungssensitivsten Items zu wählen.

[2] Zum RC-Begriff siehe Kapitel 2 auf Seite 5.

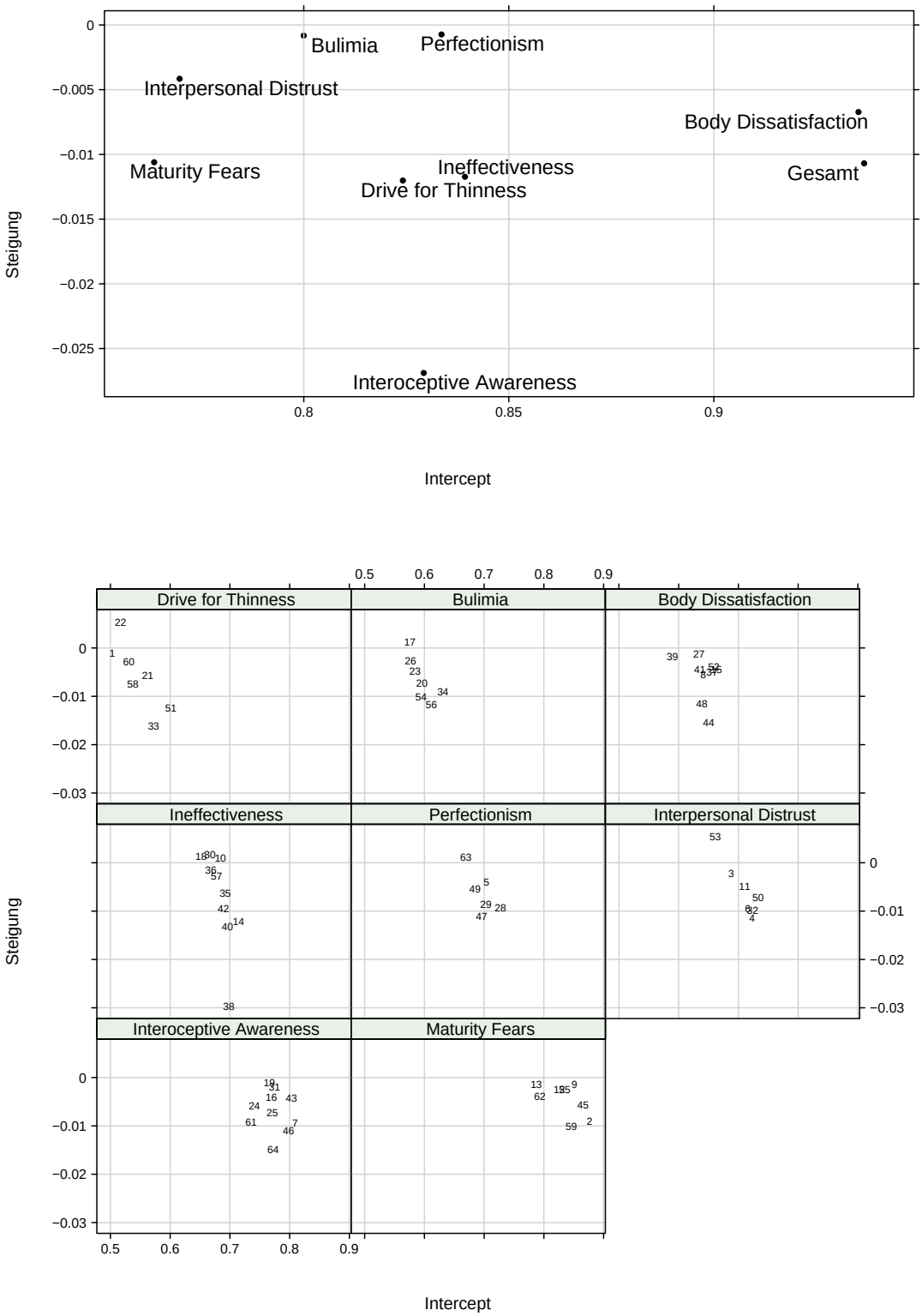

Abbildung 8.5: Retestreliabilität (Intercept) und Veränderung (Steigung) der einzelnen Skalen/Items

klassisch vs. weighted: Hiermit ist die Art der RC-Berechnung gemeint. Nach dem ‚weighted'-Ansatz (siehe Abschnitt 2.3.3 auf Seite 21) wird die Schätzung der wahren Veränderung korregiert, nach der klassischen Berechnungsmethode verändert sich der RC-Bereich.

mit Zeit: bezeichnet die hier eingeführte Methodik der zeitlichen Korrektur. D. h. es wird eine zeitliche Komponente berücksichtigt. Hierbei wird in die Originalformel des RC-Ansatz statt der Retestkorrelationen die Korrelationen abhängig pro Zeiteinheit eingesetzt.

mit Intercept: Diese Berechnung stützt sich auf die Schätzung der Reliabilität aus den oben berechneten Regressionsmodellen. Dabei wird darauf zurückgegriffen, dass der Intercept den Test-Retest Korrelationen zum Punkt Null der Zeitlinie entspricht.

mit Cronbachs Alpha: Diese Berechnung stützt sich auf die Reliabilitsberechnung nach Cronbachs Alpha (interne Konsistenz). Hierzu wird der erste Messzeitpunkt verwendet.

mit Retestreliabilität: Hierzu wurden die beiden ersten Messzeitpunkte verwandt. D.h. hierbei handelt es sich um die, neben der internen Konsistenz, geläufigsten Schätzung der Reliabilität.

Die Tabelle zeigt recht deutlich Unterschiede der geschätzten reliablen Verbesserungen. So variieren diese etwa bei der Gesamtskala zwischen 27% und 38% bei der klassischen Berechnungsweise; bei der ‚weighted' Berechungsweise zwischen 44% und 52%. Generell sind die geschätzten Verbesserungsquoten bei der klassischen Berechnungsweise niedriger als bei dem ‚weighted' Ansatz. Die höchste Veränderungsquote zeigt sich in beiden Teiltabellen bei der Skala ‚Maturity Fears'(0.48,0.62). Die geringste Veränderungsquote bei der Skala ‚Interoceptive Awareness'(0.07% und 0.14%).

8.4 Vergleich mit alternativen Erfolgskriterien

Um den neu geschaffenen RC-Index beurteilen zu können, ist ein Vergleich mit anderen, äußeren Kriterien notwendig (*Hypothesenkomplex 3* siehe Abschnitt 6.3 auf Seite 55)[1]. Als Sonderfall bietet sich hier im Bereich der Essstörungen das Gewicht an. Dieses kann als ein relativ hartes Vergleichskriterium gelten. Die Tabelle 7.1.2 auf Seite 62 gibt einen Überblick über die BMI-Werte in den drei Diagnosegruppen zum Aufnahmezeitpunkt. Der BMI ist ein etabliertes Maß zur Standardisierung des Gewichtes an der Körpergröße $BMI = kg/m^2$.

Wie ersichtlich liegen in allen drei Diagnosegruppen einige Werte über dem $BMI = 17.5$ Kriterium, dass sich als ICD-10 Kriterium für eine Anorexie durchgesetzt hat. Um

[1]Die innere, begrifflich-logische und mathematische Struktur des Indizes ist inbesondere in Abschnitt 6 auf Seite 50 erläutert worden.

Anteil gebesserter Patienten bezogen auf *ungewichtete* RC-Indizes (Prozent)

Skala	Mit Zeit	Mit Intercept	mit Chron. Alpha	mit Retestr.
Gesamt	27.09	31.32	37.71	31.32
Drive for Thinness	25.87	25.87	36.98	31.38
Bulimia	28.13	32.56	32.56	32.56
Body Dissatisfaction	24.76	29.51	35.10	34.91
Ineffectiveness	8.34	8.34	8.25	8.34
Perfectionism	13.46	13.46	17.59	17.50
Interpersonal Distrust	22.77	29.52	33.30	29.52
Interoceptive Awareness	9.16	9.63	9.63	7.38
Maturity Fears	46.00	51.37	52.97	47.60

Anteil gebesserter Patienten bezogen auf *gewichtete* RC-Indizes (Prozent)

Skala	Mit Zeit	Mit Intercept	mit Chron. Alpha	mit Retestr.
Gesamt	44.37	44.46	52.21	44.46
Drive for Thinness	48.44	48.44	48.44	48.44
Bulimia	44.81	44.99	50.40	44.99
Body Dissatisfaction	44.31	46.30	52.56	46.49
Ineffectiveness	28.39	28.39	19.43	28.39
Perfectionism	27.65	28.73	28.73	28.73
Interpersonal Distrust	38.43	46.26	46.26	46.26
Interoceptive Awareness	23.46	27.85	27.85	14.21
Maturity Fears	60.39	64.50	64.50	62.33

Tabelle 8.2: Reliable Verbesserungen nach verschiedenen Berechungsmethoden.

Gruppe [BMI]	(9.82,13.5]	(13.5,14.5]	(14.5,15.4]	(15.4,16]
Anorexia	70	69	55	48
Bulimia	0	1	6	6
Doppeldiagnose	9	11	18	25
Behandlungsdauer [Wochen]	13 ± 8.13	12.96 ± 8.75	12.09 ± 8.26	13.41 ± 7.08

Tabelle 8.3: Anzahl der verbliebenen Personen

die Personengruppe zu identifizieren, bei der im besonderen Maß das Gewicht im Vordergrund stand, wurde ein BMI-Wert kleiner als 16 bei Aufnahme in die Klinik gewählt. Hierbei kann sichergestellt werden, dass das Gewichtskriterium ein vorrangiges Behandlungsziel darstellt.

Tabelle 8.4 zeigt die Anzahl der verbliebenen Personen, wenn nur Fälle ausgewählt werden, deren BMI am Anfang der Behandlung kleiner als 16 ist. Abgetragen wurde in den Zeilen die Anzahl der Personen in den drei Diagnosegruppen sowie die Behandlungsdauer (Mittelwert und Standardabweichung). Die Einteilung der BMI in den Spalten orientiert sich in einer Quantile-Aufteilung in der Gesamtgruppe. Auch eine Regressionsanalyse zeigt, dass die Null-Hypothese, dass die Behandlungsdauer sich nicht in den drei Gruppen unterscheidet, nicht verworfen werden kann (Anova, F-statistic: 0.24, DF= 1 & 317, p-value: 0.6227).

Vor dem eigentlichen Vergleich der Kennwerte ist eine generelle Analyse des Zusammenhangs zwischen der EDI-Gesamtskala und dem BMI durchzuführen. Diese ermöglicht eine bessere Einschätzung der aus dem EDI abgeleiteten Indizes mit den auf dem BMI-basierenden Kriterien. Abbildung 8.6 auf der nächsten Seite zeigt diesen Vergleich. In der oberen Hälfte der Abbildung sind die Korrelationen wiedergegeben, in der unteren Hälfte die dazugehörigen Scatterplots mit eingetragener (lokaler) Regressionslinie. Abgetragen sind der EDI-Gesamtwert zu Aufnahme und Entlassung (‚EDI Aufn.‘,‚EDI Entl.‘), der BMI bei Aufnahme und Entlassung (‚BMI Aufn.‘, ‚BMI Entl.‘) und die Differenz des BMI und EDI zwischen Aufnahme und Entlassung (‚Diff. EDI‘, ‚Diff. BMI‘) Die Zusammenhänge zwischen dem BMI und dem EDI sind eher gering: So ist die Korrelation zwischen dem EDI und dem BMI bei Aufnahme 0.094, bei Entlassung 0.10. Auch die Differenzen korrelieren nur gering miteinander $r = -0.095$.

Damit die RC-Indices mit alternativen Erfolgsmaßen verglichen werden können, gilt es nun, alternative Kriterien zu definieren, bzw. die schon etablierten Maße vorzustellen. Die offiziellen Kriterien der Multizentrischen Essstörungsstudie (Kächele, 2000) können hierbei als schon etablierte Maße gezählt werden. Diese lauteten wie folgt[1]. Ausgegangen wurde von den vier diagnostischen Kriterien der Diagnose ‚Anorexia‘: (1) das Gewichtskriterium $BMI <= 17.5$ (2) Furcht vor Gewichtszunahme sowie (3) Störung der Körperwahrnehmung. Das vierte im DSM-III-R genannte Kriterium der Anorexie, die Auswirkungen auf den Organismus, also bei Frauen vor allem das Aus-

[1] Diese Kriterien wurden vom Autor im Zusammenarbeit mit den Mitwirkenden der multizentrischen Essstörungsstudie erarbeitet.

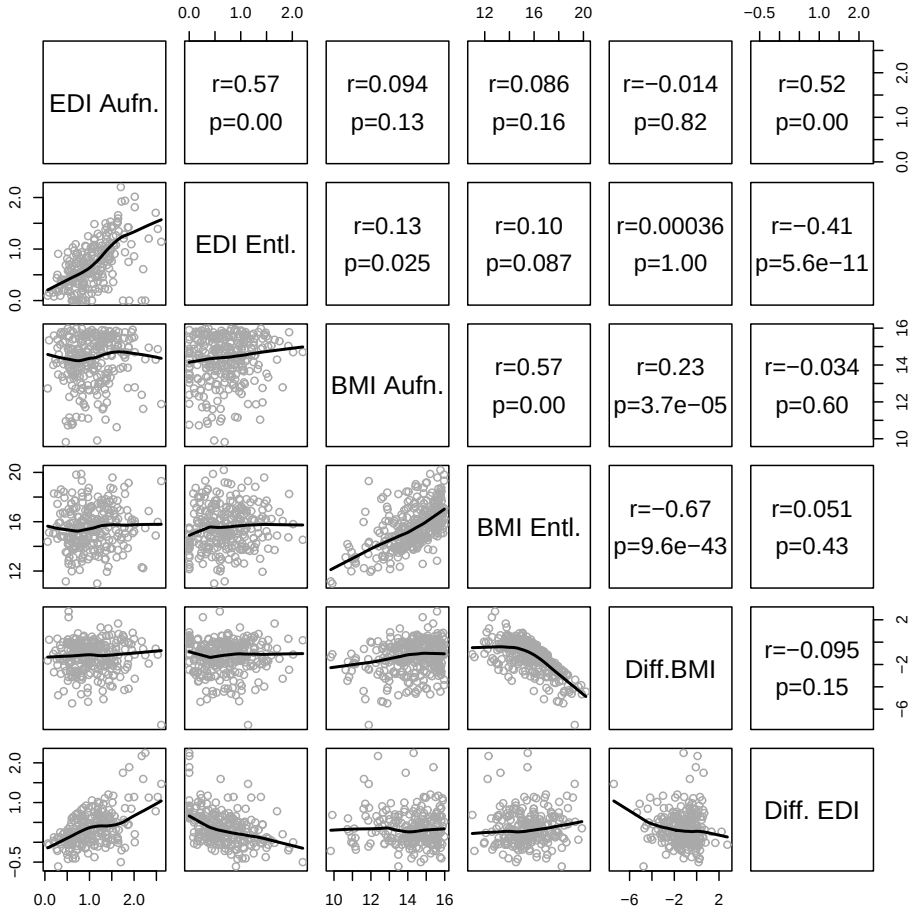

Abbildung 8.6: Vergleich des EDI und des BMI.

		Aufnahme			Entlassung		
		Ano	Bul	BulAno	Ano	Bul	BulAno
Patient	Gesundung	0	0	0	5	1	2
	Normalisierung	0	0	0	8	0	3
	Diagnosebezogener Erfolg	99	3	19	125	5	21
	kein Erfolg	144	10	44	105	7	37
Therapeut	Gesundung	0	0	0	10	1	3
	Normalisierung	0	0	0	4	1	1
	Diagnosebezogener Erfolg	67	3	15	103	3	25
	kein Erfolg	176	10	48	126	8	34

Tabelle 8.4: Das Erfolgskriterium des MZ-ESS-Studie.

bleiben der Menses, wurde nicht in die Definition des Erfolges miteinbezogen. Im Folgenden werden die vier möglichen Ausprägungen des MZ-ESS Kriteriums für Anorexie dargestellt.

Gesundung Alle drei oben genannten Kriterien liegen nach der Therapie nicht mehr im pathologischen Bereich. Dies bedeutet bzgl. der Patientenbeurteilung: Die EDI-Skala ‚drive for thinness' nimmt einen Wert ≤ 4.5 an, der BMI ist ≤ 17.5, die Differenz der Gewichtseinschätzung und das wahre Gewicht ist nicht größer als zwei BMI-Punkte. Dies bedeutet bzgl. der Therapeutenbeurteilung: Der Therapeut beurteilt die Körperschemastörung und die Angst vor Gewichtszunahme als ‚nicht vorhanden' oder ‚leicht' und der BMI ist ≤ 17.5.

Normalisierung Das Gewichtskriterium und mindestens eines der beiden psychischen Kriterien liegen nicht im pathologischen Bereich. Dies ist inbesondere von Bedeutung, da bekannt ist, dass die Störung der Körperwahrnehmung sich erst als letztes Merkmal ändert, teilweise erst lange nach einem Klinikaufenthalt.

Diagnosebezogener Erfolg Als diagnosebezogener Erfolg gilt, wenn das Vollbild dieser Störung nach der Therapie nicht mehr vorliegt. D.h. dass auf mindestens einem der oben stehenden Kriterien kein Krankheitswert mehr zu erkennen ist.

Kein Erfolg: Alle drei Bereiche liegen im pathologischen Bereich.

Tabelle 8.4 gibt den Erfolg, wie er in der MZ-ESS-Studie definiert wurde, in der oben definierten Substichprobe wieder. Wiedergegeben ist die Anzahl der Personen, die den jeweiligen Kategorien zugeordnet wurden. Da in der MZ-ESS-Studie das Erfolgskriterium jeweils durch die Patienten- und Therapeutenangaben bestimmt wurde, wurden hier auch aus beiden Quellen stammende Erfolgskriterien wiedergegeben (‚Patient' ‚Therapeut'). Wie ersichtlich, unterscheidet sich die Konzeptualisierung, je nach Quelle.

In der oben beschriebenen Konzeptualisierung handelt es sich um eine eher statische, d.h. sie enthält keine Information über das Ausmaß der Veränderung. Um ein

dem RC-Kriterium analoges Erfolgskriterium zu konstruieren, fehlt jedoch die Angabe zur Reliabilität des Messinstruments[1]. Im folgenden wurden folgende, den Aspekt der Veränderung berücksichtigende Erfolgskriterien definiert:

1. **Inkrement des BMI pro Messzeitpunkt [Diff]:** Hierbei werden der BMI zur Aufnahme vom BMI zur Entlassung abgezogen. Die vergangene (Behandlungs-) Zeit wird nicht berücksichtigt.

2. **Inkrement des BMI pro Zeit [Diff/Zeit]:** Hier wird das Inkrement des BMI zwischen Aufnahme und Entlassung pro Zeiteinheit (Behandlungswoche) betrachtet.

3. Im Gegensatz zum vorigen Punkt wird die Differenz des BMI zwischen Aufnahme und Entlassung durch die Anzahl der in dieser Zeit eingesetzten Therapiemenge geteilt. Diese teilt sich wie folgt auf (Tabelle A.3 auf Seite 113 im Kapitel A listet alle genannten therapeutischen Interventionen auf):

   a) **Therapeutische Maßnahmen [Diff/Ther.Menge]:** Hier wird nur die zeitliche Summe der therapeutischen Einzelgespräche und Gruppen betrachtet. Dabei wird im Fall der therapeutischen Gruppen durch die Gruppengröße devidiert. Ist beispielsweise eine Person eine Stunde in einer Gruppe mit insgesamt 7 anderen Patienten, wird diese Stunde als $\frac{1}{8}$ berechnet.

   b) **Gesamtmenge [Diff/Gesamt-Menge]:** Bei diesem Index wird zusätzlich zu den eigentlich therapeutischen Maßnahmen die sogenannten ‚unterstützenden Maßnahmen', also etwa die Gestaltungstherapie berücksichtigt.

Das erste Kriterium berücksichtigt, da es eine einfache Differenz zwischen den Messzeitpunkten ist, die Zeit nicht. Die zwei folgenden Kriterien, wobei das letzte eigentlich zwei bezeichnet, berücksichtigen die zeitliche Dimension, etwa als wirkliche Zeit oder als therapeutische Zeit (Stunden in Behandlung). Mit den zwei weiter oben genannten statischen Erfolgskonzeptualisierungen der MZ-ESS (MZ-ESS Pat und MZ-ESS Ther) ergeben sich also insgesamt sechs Kriterien. Um die Anzahl der Vergleiche mit den RC-Kriterien gering und damit überschaubar zu halten, werden bei den RC-Kriterien allein die sich auf die Gesamtskala beziehende Indizes berichtet. Weiterhin wurde, um die unterschiedlichsten alternativen Kriterien (also etwa die MZ-ESS-Kriterien) diese in ihrer Skalierung beibehalten. Ansonsten wäre es notwendig diese konstinuierlichen Kriterien, also etwa das Inkrement des BMI pro Zeit, zu dichotomisieren. Der Vorteil dieser Dichotomisierung läge in der Darstellbarkeit. So könnte ein Prozentwert der Übereinstimmungen angegeben werden. Der große Nachteil wäre jedoch die zwangsläufig miteinbezogene Willkür der Dichotomisierung und damit auch ein Verlust der Vergleichbarkeit. Somit werden die dreistufigen RC-Kriterien (Verschlechterung/Keine Änderung/Verbesserung) mit kontinuierlichen/rankbasierten Alternativkriterien verglichen.

Tabelle 8.5 auf der nächsten Seite zeigt die Ergebnisse der Zusammenhangsanalyse zwischen den RC-Indizes (RCI) und den alternativen Erfolgsmaßen (AM) (Spearmans

[1] Die kann nicht einfach als 1 angenommen werden, siehe auch den Abschnitt ‚Diskussion' auf Seite 9.

	Alternatives Kriterium		
RC-Index	MZ-ESS Pat.	MZ-ESS Ther.	Diff
mit Zeit	-0.004 p=0.95	-0.072 p=0.27	0.037 p=0.57
mit Zeit (w)	-0.006 p=0.93	-0.086 p=0.19	0.111 p=0.089
mit Intercept	-0.015 p=0.82	-0.092 p=0.16	0.129 p=0.047
mit Intercept (w)	-0.05 p=0.45	-0.084 p=0.2	0.134 p=0.039
mit Chron. Alpha	-0.019 p=0.77	-0.099 p=0.13	0.125 p=0.055
mit Chron. Alpha (w)	-0.05 p=0.45	-0.084 p=0.2	0.134 p=0.039
mit Retestr.	-0.017 p=0.8	-0.074 p=0.26	0.107 p=0.099
mit Retestr. (w)	-0.019 p=0.77	-0.089 p=0.17	0.121 p=0.064
RC-Index	Diff/Zeit	Diff/Ther.Menge	Diff/Gesamt-Menge
mit Zeit	0.088 p=0.18	0.01 p=0.88	0.045 p=0.49
mit Zeit (w)	0.095 p=0.14	0.048 p=0.47	0.092 p=0.16
mit Intercept	0.106 p=0.1	0.074 p=0.26	0.129 p=0.047
mit Intercept (w)	0.119 p=0.067	0.06 p=0.36	0.109 p=0.095
mit Chron. Alpha	0.107 p=0.1	0.068 p=0.3	0.119 p=0.068
mit Chron. Alpha (w)	0.119 p=0.067	0.06 p=0.36	0.109 p=0.095
mit Retestr.	0.091 p=0.16	0.058 p=0.38	0.119 p=0.067
mit Retestr. (w)	0.082 p=0.21	0.048 p=0.47	0.091 p=0.16

Tabelle 8.5: Vergleich der RC-Indizes mit verschiedenen anderen Indizes (nur Patienten mit einem $BMI < 16$ bei Aufnahme)

ρ). In den Zeilen sind hierbei die RC-Indizes, in den Spalten die alternativ formulierten Maße eingetragen. Dabei sind die (in den Zeilen stehenden) nach der gewichteten Berechnungsweise gebildeten RCI mit einem eingeklammerten w gekennzeichnet. Die Zellen repräsentieren die Spearman-Rankkorrelationskoeffizienten mit den zugehörigen P-Werten, die sich darauf beziehen, ob die Nullhypothese, dass kein Zusammenhang besteht, verworfen werden kann. Insgesamt fallen die Korrelationen zwischen den Indizes eher gering aus (Range von 0.0041 bis 0.13).

Die höchsten Korrelationskoeffizienten der auf dem RC-Konzept basierenden Erfolgsmaße waren folgende:

- Der die Zeit berücksichtigende RC-Index steht am höchsten mit der Differenz des BMI pro Zeiteinheit im Zusammenhang ($r = 0.088$),
- der die Zeit in gewichteter Weise berücksichtigende Koeffizient mit der einfachen Differenz des BMI zwischen der Aufnahme und Entlassung ($r = 0.11$),
- die RC-Indizes, die die Reliabilität durch den Intercept in ihrer gewichteten und nicht-gewichteten Version und ohne eine zeitliche Komponente bestimmen, korrelieren am stärksten mit der einfachen Differenz des BMI zwischen Aufnahme und Entlassung ($r = 0.129$ und gewichtet $r = 0.134$),
- der auf Cronbachs-Alpha basierende RC-Index, in seiner gewichteten und nicht-gewichteten Form (ohne eine zeitliche Komponente), korreliert am stärksten mit der Differenz des BMI zwischen Aufnahme und Entlassung ($r = 0.125$ und gewichtet $r = 0.134$),
- der auf der Test-Retestkorrelation beruhende Index hat den höchsten Zusammenhang mit dem Inkrement des BMI pro Gesamtmenge der therapeutischen Maßnahmen($r = 0.119$),
- in seiner gewichteten Form mit der Differenz des BMI zwischen Aufnahme und Entlassung am höchsten ($r = 0.121$).

Insgesamt sind die Korrelationen zwischen den Indizes eher klein und dementsprechend auch die Unterschiede der korrelativen Zusammenhänge.

9 Diskussion

Der folgende Abschnitt gliedert sich, analog zum Aufbau der vorliegenden Arbeit, in verschiedene, im Abstraktionsniveau unterschiedene Abschnitte. Begonnen wird mit der Diskussion der oben formulierten Hypothesen (siehe 6.3 auf Seite 55), in einem zweiten Schritt werden diese Ergebnisse zusammenfassend diskutiert und in dem letzten, dritten Schritt auf dem Hintergrund eines weiter gefassten Horizonts betrachtet. Folglich kehren die gefundenen Ergebnisse auf drei Ebenen wieder: einmal eng begrenzt in ihrer methodischen Formulierung, dann, abstrahiert von den Einzelergebnissen auf einer integrierenden methodischen Ebene, sowie als letzten Schritt aufgehoben in einen in einem weiteren, forschungslogischen Zusammenhang. Auf jeder Ebene wird versucht, Implikationen und Ausblicke für die Praxis zu geben.

9.1 Diskussion der Hypothesen

9.1.1 Hypothesenkomplex 1: Veränderungsmuster

Im Mittelpunkt stehen in diesem Hypothesenkomplex Überlegungen zur Struktur der Änderungen. Insbesondere basieren die zugrunde liegenden Annahmen zu einem modifizierten RC Ansatz auf diesen, bestimmte Strukturen voraussetzenden, Überlegungen. Im folgenden sollen demzufolge die in Abschnitt 6.3 auf Seite 55 formulierten Hypothesen auf dem Hintergrund der berichteten Ergebnisse (siehe Abschnitt 8 auf Seite 65) hinsichtlich der Frage, ob diese weiterhin gelten können[1], betrachtet werden. Insgesamt handelt es sich um die Hypothesenkomplexe 1.1 und 1.2, wobei der letztere unmittelbar auf die Veränderungsmodelle bezogen ist. So ist der Verlauf der Korrelationen unmittelbares Resultat und damit auch Indikator der Veränderungsstrukturen und wird wegen dieser inhaltlichen Bindung auch an dieser Stelle behandelt.

Insgesamt haben sich die Vermutungen zu der Richtung der Änderungen (Hypothesenkomplex 1.1) bestätigt: In der Normalstichprobe wurde keine generelle Änderungstendenz gefunden (Hypothese 1), in der klinischen Stichprobe war diese, im Sinne einer Gesundung, gerichtet (Hypothese 2). Auch hinsichtlich der Variabilität der Änderungswerte fand sich ein Unterschied (Hypothese 3). In der klinischen Stichprobe war diese höher als in der Normalstichprobe. Diese Ergebnisse sind auf dem Hintergrund des bekannten Wissens um die Stichproben und den Effekten von Psychotherapie leicht deutbar: Die klinische Stichprobe wird behandelt, d.h. die generelle Änderungstendenz

[1] Wie schon in Abschnitt 3.1 auf Seite 26 beschrieben, können Hypothesen nicht verifiziert, sondern nur nicht-falsifiziert werden.

sollte sich auch in Richtung Gesundung bewegen, während sich die Normalstichprobe, da unbehandelt (und keinem anderen systematischen Einfluss ausgesetzt), nicht verändern sollte. Auch der Unterschied der Variabilität der Änderung ist auf diesen Effekt rückführbar: Es ist bekannt, dass andere, moderierende Faktoren die Geschwindigkeit (Änderung pro Therapiemenge und Richtung (es gab auch Verschlechterungen) der Änderung determinieren.

Die Hypothesen zum Verlauf der Korrelationen (Hypothesenkomplex 1.2) können in der formulierten Weise bestehen bleiben. Wie erwartet, fallen diese mit zunehmender Zeitspanne zwischen den Messzeitpunkten (Hypothese 2, siehe Abbildung 8.4 auf Seite 69). Der Ausgangspunkt dieser Kurven ist den Reliabilitätsschätzungen durch Test-Retest und Chronbachs Alpha (interne Konsistenz) hinreichend ähnlich (Hypothese 1, siehe Tabelle 8.1 auf Seite 70). Die Differenzen zwischen beiden können hierbei als zufällige Abweichungen interpretiert werden. Da die Korrelationskurven mit zunehmender Zeitspanne zwischen den Messzeitpunkten fallen, d.h. der Betrag der Korrelationen sich immer mehr vom anfänglichen Niveau unterscheidet, vergrößert sich mithin mit zunehmender, vergehender Zeit die Differenz zu den anfänglichen Korrelationen und damit also auch zu den Annahmen der klassischen Testtheorie zur Schätzung der Reliabilität via Test-Retest-Methode. Die Form der Kurven entsprach nicht den Hypothesen (Hypothese 3). Angenommen wurde eine asymptotische (d.h. nicht-lineare) Annäherung an die Null-Korrelation. Diese Annahme begründet sich durch die Überlegung, dass die Veränderungen über die Personen hinweg unsystematisch vonstatten gingen, würde sich etwa die Reihenfolge der Personen mit der Zeit umkehren, d.h. die Personen, die am Anfang die höchsten Werte hätten, später die niedrigsten hätten (und umgekehrt), würde sich die Kurve von der positiven zur negativen Korrelation ändern, also etwa von $r = 0.8$ zu $r = -0.8$. Diese Frage ist aus dem vorliegenden Datenmaterial jedoch unentscheidbar. Beobachtet wurde in dem eingeschränkten Messintervall ein linearer Abfall (siehe Abbildung 8.4 auf Seite 69). Dabei ist ergänzend anzumerken, dass auch bei einer asymptotisch fallenden Kurve ein Teil der Kurve gut durch eine lineare Approximation darstellbar ist. Es kann hier angenommen werden, dass das Beobachtungsintervall zu kurz war und daher nur der anfängliche, in diesem Intervall gut durch eine lineare Funktion beschreibbare Teil einer Kurve beobachtet wurde. Diese Möglichkeit kündigte sich bereits in der Abbildung 6.3 auf Seite 54 an. Bei dieser ist nur bei unteren Kurven der asymptotische Verlauf sichtbar. Um hier den asymptoschen Charakter auch der flach verlaufenden Kurven zu zeigen, hätte schon bei dieser kleineren Simulation das Beobachtungsintervall größer sein müssen. Wie erwartet, fand sich hinsichtlich der verschiedenen Merkmalsbereiche ebenfalls eine merkliche Variation der Kurvenverläufe. D.h. die verschiedenen Skalen des Messinstruments verhalten sich, hinsichtlich des in der Zeit liegenden Zusammenhangs mit sich selbst, unterschiedlich (Hypothese 4 und 5). In der Skala ‚Interorecptive Awareness' fanden sich am meisten Änderungen pro Zeiteinheit, ‚Perfectionism' und ‚Bulimia ' zeigten am wenigsten Veränderungen.

9.1.2 Hypothesenkomplex 2: Der modifizierte RC-Ansatz

Die Berücksichtigung der zeitlichen Komponente verringert den prozentualen Erfolg (Hypothese 1). Hierbei ist weniger bemerkenswert, *dass* sich der Erfolg vermindert, sondern dass dieser sich nur in so einem geringen Ausmaß verringert (siehe Tabelle 8.3 auf Seite 74). Mit der gewöhnlichen Test-Retest Korrelation bestimmten RC-Index wird wird eine Rate von 31%, mit der zeitlichen Komponente werden 27.9% klinisch signifikante Verbesserungen erreicht. Ein bedeutsamer Unterschied fand sich bzgl. der internen Konsistenz. Da hier die Koeffizienten ungewöhnlich hoch waren (siehe Tabelle A.2 auf Seite 106), wurden auch mehr reliable Verbesserungen, da ja schon kleinere Veränderungen als zuverlässig gelten konnten, beobachtet. So fand sich hinsichtlich der Gesamtskala, unter Berücksichtigung von Chronbachs Alpha eine 37% Verbesserungsrate dagegen mit der (zeitlich berechneten) Retestreliabilität eine Verbesserungsrate von 31%. Die hier gefundene interne Konsistenz lag auch über der von Thiel & Paul (1988) (siehe Abschnitt 7.2 auf Seite 61) berichteten. Inhaltlich ist dies allerdings nur schwer begründbar, obwohl es auf dem Hintergrund des in der Zeit stattfindenden, fallenden Zusammenhangs zwischen den Korrelationen bemerkenswert ist, dass die quasi zeitlose Schätzung der Reliabilität durch Chronbachs Alpha durchgängig höher als die Test-Retestkorrelation ist, eben auch zu der kleinsten möglichen zeitlichen Differenz. Insgesamt kann hier festgestellt werden, dass sich die Vorstellung, dass sich bei zunehmender Zeitspanne zwischen den Messungen die Zahl bedeutsamer Änderungen im Kollektiv vorfinden lässt, also dass es eben nicht nur zufällige Schwankungen um den wahren Wert eines Individuums vorkommen, bestätigt werden.

9.1.3 Hypothesenkomplex 3: Validierung der RC-Indizes durch andere Erfolgsmaße

Um weiter der Bedeutung dieses modifizierten Ansatzes nachzugehen, wurde ein Validierungsschritt durchgeführt. Hierbei wurden Beziehungen des neu geschaffenen Indizes mit anderen, alternativen Indizes hergestellt. Da das RC-Konzept auf dem Konzept des Messfehlers basiert, war es insbesondere interessant, ein alternatives Kriterium einzubeziehen, für das der Begriff des Messfehlers nicht in dem Ausmaß zutrifft. Dies ist im Bereich der Essstörungsforschung das Gewicht, oder um präziser zu sein, das relative Gewicht pro Körpergröße, dem Body-Mass-Index (BMI)[1]. Es kann hier angenommen werden, dass es sich um eine relativ fehlerfreie Messung handelt, obwohl genaue Daten dazu nicht vorliegen. Mögliche Fehlerquellen sind dabei weniger ungenau arbeitende Waagen, als die Tendenz vieler essgestörter junger Frauen, die oft per Vertrag vereinbarte Gewichtszunahme durch bestimmte Kniffe zu umgehen, so etwa das Trinken von mehreren Litern Wasser vor dem Wiegen. Da hierzu aus dem vorliegenden Datenmaterial keine Angaben gemacht werden können und argumentiert werden kann, dass eine massive Manipulation, in auf die Behandlung von Essstörungen spezialisierte Kliniken eigentlich unmöglich sein sollte, wird die Gewichtsmessung als eine relativ messfehler-

[1] Dieser ist als das Gewicht durch die quadrierte Körpergröße $BMI = kg/m^2$ definiert. Eine eingehende Beschreibung und Kritik findet sich bei Oehlschlägel-Akiyoshi et al. (1999).

freie Messung angesehen. Um diesen Vergleich anstellen zu können, musste allerdings die Stichprobe auf die untergewichtigen Frauen beschränkt werden, da bei den anderen das Gewichtskriterium nicht in diesem Ausmaß im Mittelpunkt der Behandlung stand. Ein weiteres Außenkriterium bildete das offizielle Erfolgskriterium der MZ-ESS Studie. Dieses orientierte sich an der DSM-III-R Diagnose, somit war auch hier das Gewicht neben zwei anderen Kriterien Bestandteil an der DSM-III-R Diagnose Anorexia nervosa.

Die in den Hypothesen dargestellten Annahmen können teilweise weiterbestehen. Eine erste Orientierung bestätigt die Annahme (Hypothese 1, siehe Abbildung 8.6 auf Seite 76), dass zwischen dem EDI und dem Gewichtskriterium und dem EDI ein geringer Zusammenhang besteht. Dementsprechend fällt auch der Zusammenhang mit den RC-Indizes und den alternativen Indizes moderat aus (Hypothese 2, siehe Tabelle 8.5 auf Seite 79). Der Hauptpunkt der Argumentation jedoch, dass die zeitlich definierten RC-Indizes (die auf dem EDI beruhen) eher im Zusammenhang mit zeitlich gewichteten BMI-Verläufen, also etwa dem Inkrement des BMI pro Behandlungswochen, in Zusammenhang stehen als mit einfachen, die zeitliche Relativierung nicht enthaltende Indizes (Hypothese 3), kann nicht bestehen bleiben (siehe ebenfalls Tabelle 8.5 auf Seite 79). So ist die Korrelation des zeitlich relativierten RC-Indizes mit der Differenz des BMI zwischen Aufnahme und Entlassung, relativiert an den Behandlungswochen $r = 0.088$, und damit nur geringfügig höher als die Korrelation mit dem zeitunabhängigen MZ-ESS Kriterium für Anorexie $r = 0.04$ (Patientenbeurteilung) bzw. 0.07 (Therapeutenbeurteilung). Das geringe absolute Niveau dieser Korrelationen liegt vor allem an dem nur geringen Zusammenhang des Messinstrumentes mit dem BMI (obwohl eine Skala des EDI bei der Patientenbeurteilung für ein Unterkriterium hinzugezogen wurde). Der nicht existente Unterschied der Korrelationen widerspricht jedoch der Hypothese. Erklärt werden kann dies alleine dadurch, dass es sich um zwei verschiedene (oder nur partiell überlappende) Merkmalsbereiche, die unterschiedlichen Veränderungsprozessen unterliegen, handelt: Die zeitliche Änderung des RC-Indizes erstreckt sich in eine anderen Zeitspanne (hat eine andere Geschwindigkeit) als die Gewichtszunahme des BMI in einer Behandlung. Die Validierung liegt also hier nahe, dass es etwas anderes ist, die zeitliche Komponente in einem Merkmalsbereich zu berücksichtigen und sie dann in einem anderen einzusetzen. Dies scheint insbesondere auch der Fall zu sein, wenn diese Merkmalsbereiche inhaltlich eng verknüpft sind (essstörungsrelevante Einstellungen und der Gewichtsverlauf).

9.1.4 Zusammenfassung

Insgesamt stützen die Ergebnisse die Vorstellungen zu dem vorgeschlagenen Veränderungsmodell und den daraus abgeleiteten Konsequenzen. Vom Kostengesichtspunkt (Kosten für einen Forscher, der sein Behandlungsprogramm evaluieren möchte) ist die eingeführte Korrektur eher minimal, die Erfolgsquoten der Behandlungen verringern sich nur unwesentlich. Der Benefit inhaltlicher Schlüssigkeit wird also nur durch einen geringen Preis erbracht. Für den vorliegenden Fall muss jedoch einschränkend bemerkt werden: für einen doch recht eingeschränkten Beobachtungszeitraum. Leider war es

durch die eingeschränkte Dauer der Datenerhebung der Normalstichprobe nicht möglich, auch den Katamnesezeitraum der MZ-ESS (2.5 Jahre nach Aufnahme) zu berücksichtigen. Hier lassen sich einige gedankliche Experimente anschliessen: Angenommen die unbewiesene Annahme, die Test-Retest Korrelation fällt mit Vergrößerung des zeitlichen Abstandes asymptotisch gegen Null, dann wäre, wenn diese Korrelation sehr gering wäre, keine positive Änderung mehr möglich! Hinsichtlich mittelfristiger Beobachtungszeiträume, also etwa mehrerer Jahre, käme es mithin zu einer starken Verringerung der Erfolge (und auch der Misserfolge, wenn man keine Änderung nicht als Misserfolg rechnet). Begründet werden müsste dies etwa in dem Sinne, dass in dem Beobachtungszeitraum der Studie xy genauso viele Änderungen normalerweise stattfinden, als durch die therapeutische Maßnahme induziert werden, daher der therapeutischen Maßnahme keine Überlegenheit gegenüber einer nicht-Behandlung zuzusprechen wäre. Dem wäre entgegenzusetzten, dass im Längsschnitt wohl einige Patienten durch eine Behandlung früher gesundeten, was sich jedoch bei einem Querschnittsindex nicht abbildet. Auch ergäbe sich das Paradox, dass vielleicht einige Anorexie-Patientinnen zwar an ihrer Erkrankung sterben, jedoch aufgrund des Verlaufs der Korrelationen nicht als Verschlechterung gewertet werden würden. Das heißt, generell geht bei einer reinen Querschnittsanalyse, die Information verloren wann und für wie lange eine Änderung stattfindet. Keller et al. (1987) haben dieses Manko durch eine zeitliche Erfolgsdefinition (mindestens n-Monate in Zustand x) kompensiert.

9.1.5 Konsequenzen

Hinsichtlich des EDI ergibt sich die Frage, welche Konsequenzen die gefundenen Ergebnisse für den untersuchten Fragebogen haben und auf einer allgemeineren Ebene, welche Auswirkungen eine Betrachtung der Korrelationen über die Zeit für die Etablierung von Instrumenten überhaupt haben können.

Bei der Betrachtung der Subskalen des EDI fällt insbesondere die Skala ‚Interoceptive Awareness' auf, bei der doch ein erhebliches Maß an Veränderung über den zeitlichen Verlauf festzustellen ist. Unter dieser Skala finden sich insbesondere essstörungsrelevante Aussagen wie ‚Ich fühle mich schon nach einer kleinen Mahlzeit aufgequollen' wie auch generelle Aussagen ‚Ich bin oft verwirrt über meine wahren Gefühle' die sich unter dem Konzept der Ich-Funktionen (Rudolf et al., 1995; Leichsenring, 1999b, 1999a; Bellak, 1973) zusammenfassen lassen. Es liegt die Annahme nahe, dass die starken Veränderungen auf dieser Skala mehr als auf anderen Skalen sowohl durch den Prozess des Älterwerdens als auch von starken situativen Aspekten geprägt sein könnten. Auf jeden Fall ist eine Skala, in der in einer nicht behandelten Stichprobe sich ein großes Maß an Veränderungen findet, kaum für die Messung einer induzierten Veränderung geeignet. Diese würde gleichsam aufgrund des starken , Rauschen' der Grundrate von Veränderung sich nicht bemerkbar machen. Bzw. wäre es nach der Logik der vorliegenden Arbeit schwer, eine therapiebedingte Änderung herbeizuführen, da der Bereich in dem keine Änderung festgestellt werden würde, sich mit verstreichender Zeit vergrößerte. Dies spricht auch gegen die verbreitete Praxis, Skalen in denen sich viel Veränderung findet, sofort Änderungssensivität zuzusprechen.

Hinsichtlich der Skalenkonstruktion *überhaupt*, kommt man zu den vielleicht etwas überraschendes Urteil, dass die Beachtung der natürlichen Änderungsfrequenz auf dem Instrument gerade so sein sollte, dass hier wenig spontane Änderungen stattfinden, da diese sonst die wahren Änderungen diskreditierten. Dies Urteil kann jedoch nicht pauschal gelten, so sind immer die Gründe für eine starke zeitliche Variabilität (Entwicklungseffekte, situative Komponenten, kulturelle Einflüsse) zu reflektieren.

9.2 Methodologische Implikationen

Im folgenden sollen mögliche methodologische und inhaltlichen Implikationen nochmals allgemeiner diskutiert werden.

Das Modell der klassischen Testtheorie ist nur für sehr kurze Zeiträume anzuwenden, in längeren Zeiträumen summieren sich die substantiellen Änderungen auf, eine Abnahme der Test-Retestkorrelationen ist zu beobachten. Dies verweist auch auf die immanente ‚Zeitlosigkeit' der klassischen Testtheorie (vgl. insbesondere Punkt ‚Operationale Definition des wahren Wertes' auf Seite 15). Als substantiell erscheinen in diesem Zusammenhang alle Änderungen, die über die zufälligen Schwankungen um einen wahren Wert hinausgehen (siehe insbesondere Abbildung 6.2 auf Seite 53). Nach einer modifizierten klassischen Testtheorie, der state-trait Theorie, (Steyer & Schmitt, 1998; Steyer, 1992, 1984; Eid, 1995) lassen sich die einzelnen Bestandteile als a) wirkliche Änderung (trait-change), b) situative (state-change) und c) Messfehler einteilen. Für den vorliegenden Zusammenhang ist es unerheblich, ob es sich bei den Schwankungen (vgl. Abbildung 6.2 auf Seite 53) um wahre situative Veränderungen oder um Messfehler handeln, da im Kontext der Psychotherapieforschung nur die wirklichen Änderungen von Interesse sind. Jedoch bleibt zu konstatieren, dass die beobachteten Verläufe der Korrelationen nur auf einen derartigen theoretischen Hintergrund verstehbar sind.

Weiterhin ist an dieser Stelle die oben genannte Schwierigkeit mit dem Ansatz, der Verminderung der Ergebnisse bei hohen Schwankungen der wahren Werte auf dem Zeitstrahl, zu diskutieren. Hierbei soll vor allem auch auf andere Konzepte eingegangen werden. Zur Erinnerung: Es handelt sich bei dem hier vorgeschlagenen Ansatz um eine Korrektur bzgl. der normalerweise zu erwartenden Änderungen. Hier besteht jedoch eine Differenz dem Konzept der Spontanheilungen, das ja mit einer gewissen Berechtigung auch als normalerweise zu erwartende Änderungen bezeichnet werden könnte, allerdings bezieht sich die Normalität eben bei dem Begriff der Spontanheilung nicht auf eine Normalstichprobe, sondern auf ein klinisches Sample. Insbesondere das oben erwähnte Paradox, dass bei längeren Behandlungszeiten fast keine Änderungen der Merkmalsausprägung als wirkliche Änderungen gezählt werden würden, macht eine Modifikation des vorgeschlagenen Ansatzes notwendig. Es gibt keinen Grund anzunehmen, dass die in der Normalstichprobe geschätzte normalerweise zu erwartende Veränderung in einer unbehandelten klinischen Stichprobe (d.h. Spontanheilungsquote) entspricht[1]. Inhaltlich würde jedoch die Einbeziehung der Spontanheilungsquote

[1]In der klassischen Testtheorie ist die problematische Annahme enthalten, dass der wahre Wert nicht mit dem Fehler im Zusammenhang steht. Dies ist jedoch oft unplausibel. So kann argumentiert werden,

mehr Sinn machen, geht es doch bei einer Evaluation der therapeutischen Veränderung immer um den Nachweis, dass diese über der Spontanheilungsquote liegt . Im welchem Verhältnis die Spontanheilungsquote zu der normalerweise beobachtbaren Änderung steht, konnte in der vorliegenden Untersuchung jedoch nicht beleuchtet werden. Die vorgeschlagene Schlussbildung würde etwa lauten: In der Studie xy wurden $x\%$ Verbesserungen, die über den in diesem Zeitraum erwarteten Spontanheilungsquote liegen, gefunden. **Zusammenfassend wird die Einführung des zeitbasierten RC-Indizes befürwortet, jedoch mit der Einschränkung, dass sich die normalerweise zu beobachtende Änderung auf die Spontanheilungsquote beziehen sollte.** Analog zu der Definition 2.4 auf Seite 22 ergibt sich folglich die folgende, neue Definition des CS-Ktriteriums:

> Unter dem Begriff CS werden diejenigen Erfolgsmaße bezeichnet, bei denen die folgenden beiden Kriterien beide erfüllt sein müssen. (1) Es muss ein Wechsel von einem krankheitswertigen zu einem nicht mehr krankheitswertigen psychischen Zustand vorliegen. (2) Muss die beobachtete Veränderung über dem in den Beobachtungszeitraum liegende spontan auftretende Gesundungsrate liegen und nicht in dem durch den Messfehler bedingten Bereich liegen.

Ein weiterer Vergleich soll diese Argumentation weiter illustrieren. Angenommen es handelte sich bei der zu behandelnden Krankheit nicht um eine Essstörung sondern um eine Erkältungskrankheit, also einen gemeinen Schnupfen. Im Gegensatz zu der potentiell chronischen Essstörungserkrankung kann bei einem Schnupfen davon ausgegangen werden, dass in ca. zehn Tagen bei einem Großteil der erkrankten Personen der Schnupfen nicht mehr besteht. Nimmt man weiter an, dass nach jedem Tag $\frac{1}{10}$ der Personen den Schnupfen verlieren und man nach fünf Tagen messen würde, würde man nach der alten Berechnungsweise richtigerweise 50% Besserungen finden. Ohne Information zu der normalerweise zu erwartenden Änderung des Krankheitszustandes könnte man fälschlicherweise annehmen, dass ein zufällig verabreichtes, jedoch wirkungsloses Medikament 50% Besserungen bewirkt hätte. Würde allerdings die Rate der Spontanheilungen nach dem hier vorgeschlagenen Index berechnet werden, wäre die Erfolgsquote auch gleich Null. Dieses Gedankenspiel zeigt auch drastisch die Folgen auf, die Variabilität eines Merkmals durch die Veränderung in der Normalstichprobe zu erfassen: Hier wären in der Tat fast nur die Neuerkrankungen berücksichtigt, dass heißt, die eigentlich relevante Information, die Spontanheilungsquote pro Zeiteinheit, fehlt.

Hier stellt sich unmittelbar die schon in Abschnitt 2 auf Seite 5 angeschnittene Frage nach der Nützlichkeit des vorgeschlagenen Indizes in den verschiedenen Zusammenhängen, in denen er möglicherweise Verwendung finden könnte. Der CS-Ansatz

dass die Extrembereiche eines Merkmalsbereichs schwerer messbar sind. In der vorliegenden Arbeit eröffnet sich ein analoger Sachverhalt. So wird auch hier angenommen, dass die Rate der Änderungen vom wahren Wert (Normalstichprobe/klinische Stichprobe) abhängt. D.h. dass die Schwankungen in der Normalstichprobe etwas anderes als die Schwankungen in der klinischen Stichprobe (Spontanheilungsrate) darstellen.

in seiner klassischen, wie auch in der modifizierten Form rechtfertigten sich vor allem durch seine semantische Bedeutsamkeit; hier liegt sein größter Vorteil. In einem sehr verdichteten Index wird die Wirksamkeit einer Behandlungsmaßnahme dargestellt. Die klinische Relevanz spiegelt sich in dieser Form hinreichend wieder. Der Vorteil dieser gezielten Verdichtung ist jedoch in einem anderen Zusammenhang auch ihr größter Nachteil: So ist ein solcher Index etwa in klinischen Arbeiten, die die Prädizierbarkeit von Ergebnissen zum Ziel haben, wegen des Wegfalls von Information von Nachteil, so wird ja bei der Umrechnung eine kontinuierliche Variable dichotomisiert. Konkret bedeutet dies, dass zur Beantwortung solcher Fragestellungen bei dem Einsetzen derartig kondensierter Indizes mehr Daten benötigt werden. Dies bedeutet, dass die Kosten die mit dem Einsatz eines derartigen Ansatzes verbunden sind, steigen. Ähnliche Überlegungen gelten auch für Metaanalysen.

9.3 Beschränkungen der vorliegenden Studie

U.a. auf dem Hintergrund der oben durchgeführten Überlegungen ist die vorliegende Studie wie folgt zu kritisieren:

1. Die Normalstichprobe sollte keine Normalstichprobe sondern eine unbehandelte Stichprobe von erkrankten Personen umfassen (siehe die Argumentation weiter oben).

2. Zur Verdeutlichung des Effekts, der mit der Einführung des Indizes verbunden ist, ist die Wahl der Krankheit (Essstörungen) ungünstig. Dies ist deshalb der Fall, da bei Essstörungen relativ wenige positive Änderungen vonstatten gehen und daher die Differenzen verschiedener Berechungsweisen gering ausfallen.

Während der erste Punkt grundsätzlicher Natur ist, deutet der zweite Punkt auf ein Erschwernis der Darstellung. Wünschenswert wäre zur besseren Darstellung der Effekte eine Erkrankung mit kürzerer Erkrankungsdauer und mit einem geringeren Anteil chronischer Verläufe gewesen. An dieser Stelle ist auch noch einmal auf die langen Erkrankungsdauern bei Beginn der Studie hingewiesen (7.3 Jahre).

9.4 Theoretische Reflexion auf der Ebene der Forschungslogik

> „Die Mathematiker sind eine Art von Franzosen: Redet man zu ihnen, so übersetzen sie es in ihre Sprache, und dann ist es alsbald etwas anderes."
> Goethe (1981)

Betrachten wir die vorhergehenden Schritte, die theoretische Revision des Begriffs, die mathematische Formulierung und die Anwendung auf empirische Daten, so stehen zwischen diesen Schritten verschiedene Prozesse, oder anders formuliert, der Forscher

sieht sich verschiedenen Aufgaben gegenüber. Diese Aufgaben könnten etwa wie folgt lauten:

Analyse der theoretischen Begrifflichkeit: Als erstes ergibt sich in der Regel aus einem begrifflichen Zusammenhang die Forderung, eine neue Begrifflichkeit zu (umzu-) definieren. Der RC-Begriff ist hier jedoch eine Ausnahme, da er als bewusste Abgrenzung zu anderen statistischen Schlussverfahren gebildet wurde (siehe Abschnitt 3 auf Seite 24)[1].

. . . an sich: Hier wurde als solches der theoretische Begriff hinsichtlich seiner inhaltlichen Schlüssigkeit aus sich selbst heraus, durch den Vergleich mit seiner semantischen Bedeutung und seiner bestehen Konzeptualisierung, analysiert. Hier zeigte sich insbesondere, dass der etablierte RC-Begriff inhaltlich nicht schlüssig ist. Anzumerken ist, dass für eine derartige Begriffsanalyse in der Psychologie technische Richtlinien fehlen.

. . . in ihrem funktionellen Zusammenhang: Diese Analyse der Begrifflichkeit fokussiert auf die Funktion, die diese in ihrem Zusammenhang (Forschung) einnimmt und beleuchtet die externale Schlüssigkeit. Begonnen wurde mit der alltagssprachlichen Relevanz: der gegenwärtige RC-Begriff ist kaum im Alltagsdiskurs verständlich, d.h. bzgl. der Funktion eines solchen Indizes, nach außen hin Aussagen zu vermitteln, erscheint der RC-Index in seiner bisherigen Formulierung als ungeeignet. Hier muss auf die Differenz zu dem vorigen Punkt ‚Analyse der theoretischen Begrifflichkeit' hingewiesen werden. Beide Punkte bezeichnen eine Analyse der Begrifflichkeit hinsichtlich seiner Rolle in einem gewissen Kontext. Dabei ist jedoch der erste eher als die Geschichte, als Anstoß derartige Überlegungen zu beginnen; der zweitere dagegen als vertiefende Überlegung zu verstehen. D.h. der Unterschied zwischen beiden Punkten besteht eher in der Abfolge der Argumentation als der Sache nach.

Übersetzung ins Mathematische Die Begrifflichkeiten stehen jedoch in der empirischen Forschung nie für sich, es tritt eine weitere Ebene hinzu, nämlich die Umsetzung in eine abstrakte mathematische Formel. Erst dadurch kann die Begrifflichkeit Zahlen und damit Ergebnisse produzieren. D.h. durch den Weg der Mathematisierung bereichert sich die Theorie um neue Fakten. Doch auch bzgl. der logischen Entwicklung eines empirischen Begriffs vollzieht sich analog dazu etwas ähnliches. Ein empirischer, methodischer Begriff muss mathematisiert werden um Anwendbar zu werden. Durch diese Umsetzung gewinnt er an Bedeutung.

Interne Konsistenz: Wie die sprachliche Analyse der Begrifflichkeit so muss auch

[1] An dieser Stelle ist nicht die immanente Negativität der Sprache gemeint, wie dies etwa von de Saussure (de Saussure, 1967) in seiner Sprachtheorie formuliert wurde. De Saussure betont den strukturellen Aspekt der Sprache. Ein Signifikant (der Bezeichnende) ist immer auch durch seinen Ort (strukturellen Platz) im Sprachgefüge gekennzeichnet, also auch durch das, was er nicht ist (Negativität).

die mathematische Formulierung in sich widerspruchsfrei sein. Hier ist insbesondere auch die Kritik von Maassen (1998, 2000b, 2000a) an den alternativen RC-Formulierungen zu nennen (siehe Kapitel 2.3.3 auf Seite 18).

Der mathematische Zusammenhang: Die mathematische Formulierung ist jedoch ebenfalls in einen Zusammenhang mathematischer Theoriebildung zu stellen. Dieser andere, mathematische Zusammenhang bildet zwar als Ganzes bestimmte Aspekte des theoretisch-semantischen Zusammenhangs ab, folgt jedoch einer eigenen Logik und trifft etwa eigene Unterscheidungen zu anderen Methodiken. Neben dieser Abgrenzung zu anderen Verfahren ist es hier insbesondere von Wichtigkeit. implizite, hinzukommende Annahmen zu reflektieren. Diese Reflektion ist immer auch eine inhaltliche. D.h. der Prozess der Mathematisierung strahlt gleichsam in die inhaltliche Ebene zurück, es kommt zu einer inhaltlichen Konkretisierung und Bereicherung (siehe Abschnitt 6 auf Seite 50). Diese Bereiche machen im besonderen Maße die Rolle des Normalitätbegriffs (siehe 4 auf Seite 34) deutlich.

Strukturelle Analyse der Übersetzung Als ‚Artefakt' dieser auf die Forschungslogik reflektierenden Arbeit wurde eine strukturelle Analyse des Übersetzungsprozesses durchgeführt (vgl. Abschnitt 6 auf Seite 50), bei der sich eine Asymmetrie des RC-Begriffs zeigte . Diese war zwar als solche nicht hinreichend, sondern musste durch eine inhaltliche Begründung ergänzt werden, war jedoch bzgl. der Klarifikation des Vorganges hilfreich.

Empirische Erprobung: Durch die konkrete Anwendung eines neuen Ansatzes eröffnet sich ein weiterer Wissensbereich, der des Anwendungswissens. Hier lassen sich verschiedene Momente unterscheiden:

Einschätzung der Voraussetzungen: Hier erscheint es notwendig, die zugrundeliegenden, mathematischen Annahmen zu reflektieren. Im vorliegenden Fall war dies die Evaluierung des zugrunde liegenden Veränderungsmodells.

Einschätzung der Auswirkungen: Im Gegensatz zu dem üblichen hypothesentestenden Schlussverfahren hat die Anwendung des neuen Ansatzes eher den Charakter eines Schätzverfahrens. Es stehen weniger die einzelnen Hypothesen als solche im Vordergrund, als die Einschätzung, welche Auswirkungen und Implikationen ein solcher Ansatz nach sich zieht. Hier hat sich gezeigt, dass die Korrektur auf einen kurzen Zeitraum wenig Auswirkungen auf die Ergebnislage hat.

Vergleich mit anderen Konzepten: Die theoretische Reflexion rekurriert auf die veränderte funktionelle Rolle eines modifizierten RC-Indizes im Forschungsgeschehen. Diese funktionelle Rolle ist in Abgrenzung zu ähnlichen Konzepten zu analysieren. Hierbei erbrachte der Vergleich mit der Spontanheilungsquote eine inhaltliche Modifikation, welche den hier vorgestellten Ansatz einer grundsätzlichen Kritik aussetzte.

Um es mit der Wissenschafttheorie Heideggers (siehe Abschnitt 5 auf Seite 42, Heidegger, 1990c, 1980) noch einmal zusammenzufassen: Im Ganzen handelt es sich bei den oben beschriebenen Schritten um einen Teilaspekt im Vorgang des Einrichtens eines bestimmten Gegenstandsbezirkes. Durch diese Einrichtung entsteht die Möglichkeit der Produktion gesicherten Wissens. Dabei wirkt das schon vorher Gewusste (oder erst Hergeleitete), das (in einem sehr allgemeinen Sinne) Mathematische, gegenstandskonstituierend und das Forschungsfeld sichernd. Diese Sicherung ist die Strenge der Forschung, in deren Produktionsprozess des Wissens. Die hier beobachtete Änderung der Methodik konstituiert jedoch nicht einen grundsätzlich neuen Gegenstand (und entwirft und umgrenzt keinen neuen Gegenstandsbereich), sondern verändert sozusagen eine Maschine im Maschinenpark aufeinander bezogener Maschinen im Produktionsprozess der Forschung. Der Prozess als solcher ist in sich reflexiv, die semantische Theoriebildung wird durch den Zerrspiegel der Mathematik auf sich selbst zurückgespiegelt.

Was ist aus dieser allgemeinen Betrachtung zu lernen? Im Sinne von Gigerenzer (2000, 1998, 1989b, 1989a), Gigerenzer & Murray (1987) kann hier die Forderung an die Forschung gestellt werden, die Methodik nicht als starres, nicht hinterfragbares Mittel zu sehen, welches halb bewusstlos, mechanisch angewandt wird, sondern als etwas, das immer auch in seiner Funktionalität reflektiert werden muss. Gigerenzer zitiert hier oft die Gründungsphase der Psychologie, in der die Methodik den Problemen entsprechend entwickelt wurde. Die vorliegende Arbeit sollte diesen dynamischen Aspekt methodischen Handelns betonen.

10 Zusammenfassung

Ziel der Arbeit ist die Einführung und Abschätzung der Auswirkungen eines modifizierten Reliable Change (RC) Indizes. Durch diese Modifikation wird das Konzept der klinischen Signifikanz (CLS) erweitert, so dass dieses neben der Frage, ob sich die Symptomatik des Patienten nach der Behandlung im Normbereich befindet, auch die Frage, ob die Veränderung des Patienten während der Therapie außerhalb der ‚normalerweise zu erwartenden Veränderung' liegt, umfasst. So konstituiert der Normalitätsbegriff sowohl den dynamischen (RC) als auch den statischen Teil des CLS-Konzepts und mithin wird eine begriffliche Asymmetrie des Konzepts ausgeglichen. Voraussetzung für eine derartige Erweiterung ist ein bestimmtes, hinreichend allgemeines Veränderungsmodell, aus dem konkrete Hypothesen ableitbar sind. *Methode*: Zur Bestimmung des normalerweise zu erwartenden Maßes der Veränderung relevanter, essstörungsspezifischer Einstellungen (EDI) wurden 295 gesunde Frauen zu insgesamt 5 Zeitpunkten untersucht. Als klinische Stichprobe wurden Daten der multizentrischen Essstörungsstudie von 1171 Frauen verwendet. Anhand der Normalstichprobe wurde das Ausmaß der normalerweise zu erwartendender Veränderung pro Zeiteinheit bestimmt und zur Bestimmung des zur Zeit konditionalen RC-Cut-Offs hinzugezogen. *Ergebnisse*: Wie erwartet, zeigt sich mit Vergrößerung des zeitlichen Abstands zwischen den Messungen eine Abnahme des Zusammenhanges zwischen den Skalen; das Ausmaß dieser variiert sehr stark zwischen den Subskalen des eingesetzten Instrumentes. Entgegen der Erwartung konnten bestimmte Folgerungen aus dem zugrundeliegenden Veränderungsmodell, so der bei Vergrößerung des zeitlichen Abstands asymptotisch gegen Null strebende Abfall der Korrelationen, nicht bestätigt werden. Bei der Schätzung des Therapieerfolgs fanden sich moderate Unterschiede zwischen den beiden RC-Indizes (klassischer Ansatz RC: 31.3%, mit Zeit RC: 27.1%). Eine Validierung des Indizes ergab nur geringe Zusammenhänge zu alternativen Erfolgskriterien, auch zu denjenigen, die ebenfalls eine zeitliche Komponente berücksichtigten. *Diskussion*: Es kann angenommen werden, dass das zugrundeliegende Veränderungsmodell weiterhin gilt. So können bestimmte, diesen widersprechende Ergebnisse, auf den zu kurzen, zeitlichen Abstand zurückgeführt werden. Die geringen Zusammenhänge zwischen den neuformulierten Index und den alternativen Kriterien sind auf die geringen Zusammenhänge des EDI mit dem relativen Gewicht zurückzuführen. Insgesamt stellt diese Neuformulierung des CLS-Konzepts eine in sich konsistente Erweiterung dar. Die Schätzung der normalerweise zu erwartenden Änderung durch die Erhebung einer Normalstichprobe führt jedoch zu Widersprüchen, so dass diese durch eine klinische Stichprobe (Spontanheilungsquote) zu erfolgen hat. So kann auch der fehlende Zusammenhang zu anderen, die zeitliche Komponente berücksichtigenden Erfolgsmaßen durch diese Inkonsistenz

erklärt werden.

Literaturverzeichnis

Abrams, K., Ashby, D., & Errington, D. (1994). Simple bayesian analysis in clinical trials: A tutorial. *Controlled Clinical Trials*, *15*.

Antonovsky, A. (1985). *Health, stress and coping.* San Francisco: Jossey-Bass.

Antonovsky, A. (1987). *Unraveling the mystery of health: How people manage stress and stay well.* San Francisco: Jossey-Bass.

Bellak, L. (1973). *Ego functions in schizophrenics, neurotics, and normals.* New York: John Wiley & Sons, Inc.

Berry, D. A. (1996). *Statistics: A basian perspective.* Belmond: Wadsworth.

Bregg, C. B. (1991). Advances in statistical methology for diagnostic medicine in the 1980's. *Statistics in Medicine*, *10*, 1887–1895.

Carnap, R. (1928). *Der logische Aufbau der Welt.* Hamburg: Felix Meiner Verlag.

Chambers, J. M. (1999). *Greater or lesser statistics: A choice for future research* (Tech. Rep.). Murray Hill, New Jersey, `http://cm.bell-labs.com/cm/ms/departments/sia/jmc`: AT&T Bell Laboratories.

Christensen, L., & Mendoza, J. L. (1986). A method of assessing change in a single subject: an alteration of the rc index. *Behavioral Therapy*, *17*(305–308).

Cleveland, W. S. (1993). *Visualizing data.* New Jersey: Hobart Press.

Cleveland, W. S. (1994). *The elements of graphing data.* New Jersey: Hobart Press.

Cook, R. J., & Farewell, V. T. (1996). Multiplicity considerations in the design and analysis of clinical trials. *Journal of the Royal Statistal Society Serie A*, *159*, 93-110.

Cowles, M. (1989). *Statistics in psychology: an historical perspective.* New Jersey: Lawrence Erlbaum Assosciates.

Danzinger, K. (1987). Statistical method and historical development of research practice in american psychology. In L. Krüger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution: Volume 2: Ideas in sciences* (pp. 35–48). Cambridge: A Bradford Book.

Danzinger, K. (1990). *Constructing the subject: Historical origins of psychological research.* Cambridge: Cambridge Unsiversity Press.

de Saussure, F. (1967). *Grundfragen der allgemeinen Sprachwissenschaft.* Berlin: Walter de Gruyter.

Ebbutt, A. F., & Frith, L. (1998). Practical issues in equivalence trials. *Statistics in Medicine*, *17*, 1691–1701.

Eid, M. (1995). *Modelle der Messung von Personen in Situationen.* Weinheim: Beltz Verlag.

Emerson, S. S. (1995). Stopping a clinical trial very early based on unplanned interim analysis: A group sequential approach. *Biometrics*, *51*(1152–1162).

Everitt, B. S. (1998). *The cambridge dictionary of statistics.* Cambridge University Press.

Eysenck, H. J. (1957). The effects of psychotherapy: An evaluation. *Journal of Consulting Psychology*, *16*, 319–324.

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh: Oliver & Boyd.

Fisher, R. A. (1935). *The design of experiments.* Edinburgh: Oliver & Boyd.

Fisher, R. A. (1956). *Statistical methods and scientific inference.* Edinburgh: Oliver & Boyd.

Formann, A. K. (1984). *Die latent-class-analysis.* Weinheim: Beltz Verlag.

Foucault, M. (1990). *Die Ordnung der Dinge.* Frankurt: Suhrkamp.

Freud, S. (2000). Eine Kindheitserinnerung des Leonardo da Vinci. In *Studienausgabe* (pp. 87–152). Frankfurt am Main: Fischer.

Gadenne, V. (1977). *Die methodischen Gundlagen psychologischer Untersuchungen.* Inaugural-Dissertation.

Gadenne, V. (1984). *Theorie und Erfahrung in der psychologischen Forschung.* Tübingen: J. C. B. Mohr (Paul Siebeck).

Garner, D. (1991). *Eating disorder inventory-2 professional manual.* Odessa, FL: Psychological Assessment Resources.

Garner, D., Garfinkel, P., & O'Shaughnessy, M. (1985). The validity of the distinction between bulimia with and without anorexia nervosa. *American Journal of Psychiatry*, *142*, 581–587.

Garner, D., & Olmsted, M. (1984). *Eating disorders inventory manual.* Odessa, FL: Psychological Assessment Resources.

Garner, D. M., Olmstedt, M. P., & Polivy, J. (1983). Development and validation of a multidimensional eating disorder inventory for anorexia nervosa and bulimia. *International Journal of Eating Disorders*, *2*, 15–34.

Gelman, A., Clarin, J. B., Stern, H. S., & Rubin, D. B. (1997). *Bayesian data analysis.* London: Chapman & Hall.

Gigerenzer, G. (1989a). Probabilistic thinking as the fight against subjectivity. In L. Krüger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution. volume 2: Ideas in sciences.* Cambridge: A Bradford Book.

Gigerenzer, G. (1989b). Survival of the fittest probabilist: Brunswik, Thurstone, and the two disciplines of psychology. In L. Krüger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution. volume 2: Ideas in sciences.* Cambridge: A Bradford Book.

Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences*, *21*(2), 199–200.

Gigerenzer, G. (2000). *Adaptive thinking.* Oxford University Press.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics.* New Jersey: Lawrence Erlbaum Associates.

Glasersfeld, E. von. (1997). *Radikaler Konstruktivismus.* Frankfurt am Main: Suhrkamp.

Goethe, J. W. (1981). *Maximen und Reflexionen.* Weimar : Aufbau-Verlag.

Hageman, W. J. J., & Arrindell, W. A. (1993). A further refinement of the reliable change (RC) index by improving the pre-post difference score: Introduction RC_{ID}. *Bevaviour Research and Therapy*, *31*(7), 693–700.

Hegel. (1832). *Wissenschaft der Logik, Band 1 Seinslogik.* Frankfurt am Main: Suhrkamp.

Heidegger, M. (1923). *Sein und Zeit.* Tübingen: Niemeyer.

Heidegger, M. (1980). Zeit des Weltbildes. In *Holzwege* (pp. 73–110). Frankfurt am Main: Vitorio Klostermann.

Heidegger, M. (1990a). Die Frage nach der Technik. In *Vorträge und Aufsätze* (pp. 9–40). Pfullingen: Neske.

Heidegger, M. (1990b). Was heißt Denken? In *Vorträge und Aufsätze* (pp. 123–137). Pfullingen: Neske.

Heidegger, M. (1990c). Wissenschaft und Besinnung. In *Vorträge und Aufsätze* (pp. 41–66). Pfullingen: Neske.

Heidegger, M. (1996). Der Satz der Identität. In *Identität und Differenz* (pp. 9–30). Pfullingen: Neske.

Heidegger, M. (1997). *Der Satz vom Grund.* Pfullingen: Neske.

Hollander, M., & Wolfe, D. A. (1973). *Nonparametric statistical inference.* New York: John Wiley & Sons, Inc.

Hsu, L. M. (1989). Reliable changes in psychotherapy: Taking into account regression toward the mean. *Behavioral Assesssment*, *11*, 459–467.

Hubbard, R., & Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology–and its future prospects. *Educational and Psychological Measurement*, *60*(5), 661–681.

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, *5*, 299—314.

Jacobson, N. S., Folette, W. C., & Revensdorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavioural Therapy*, *17*, 308–311.

Jacobson, N. S., & Revensdorf, D. (1988). Statistics for assessing the clinical significance of psychotherapy techniques: Issues, problems, and new developments. *Behavioral Assessment*, *10*, 133–145.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12–19.

Jaynes, E. T. (1984). The intuitive inadequacy of classical statistics. *Epistemologia*, *7*, 43–74.

Jaynes, E. T. (1985). Maximum entropy and bayesian methods in applied statistics. In J. H. Justice (Ed.), *Bayesian methods: General background. an introductory tutorial* (pp. 1–25). Dordrecht: Kluwer Academic Publishers.

Jaynes, E. T. (1990). Probability theory as logic. In P. F. Fougere (Ed.), *Maximum entropy and bayesian methods in applied statistics* (pp. 5–40). Dordrecht: Kluwer Academic Publishers.

Jaynes, E. T. (2003). *Probability theory: The logic of science.* Cambridge University Press.

Kächele, H. (1999). Eine multizentrische Studie zu Aufwand und Erfolg bei psychodynamischer Therapie von Essstörungen. Studiendesign und erste Ergebnisse. *Psychotherapie Psychosomatik Medizinische Psychologie*, *49*, 100–108.

Kächele, H. (2000). *Ergebnisse der multizentrischen Essstörungsstudie.* (Abschlussbericht)

Kächele, H., Kordy, H., & Richard, M. (2001). Therapy amount and outcome of inpatient psychodynamic treatment of eating disorders in germany: Data from a multicenter study. *Psychotherapy Research*, *11*, 239-257.

Keller, M. B., Lavori, P. W., Friedman, B., Nielsen, E., Endicott, J., McDonald-Scott, P., & Andreasen, N. C. (1987). The longitudinal interval follow-up evaluation. *Archieves of General Psychiatry*, *44*, 540–548.

Kluge, F. (1999). *Etymologisches Wörterbuch der deutschen Sprache.* De Gruyer.

Kordy, H. (1986). *Über den Umgang mit Beobachtungen in der Psychologie: zum Verhältnis von Beobachtungen, Modellkonstruktion und Strukturerkenntnis.* Frankfurt am Main: Europäische Hochschulschriften, Reihe 6.

Kordy, H. (1997). Das Konzept der Klinischen Signifikanz in der Psychotherapieforschung. In B. Strauß & J. Bengel (Eds.), *Forschungsmethoden in der Medizinischen Psychologie. Jahrbuch der Medizinischen Psychologie 14* (pp. 129–145). Göttingen: Hogrefe.

Kordy, H., Percevic, R., & Martinovich, Z. (2001). Norms, normality, and clinical significant change: Implications for the evaluation of treatment outcomes for eating disorders. *Journal of Eating Disorders*(76-186).

Kuhn, T. S. (1967). *Die Struktur wissenschaftlicher Revolutionen.* Frankfurt am Main: Suhrkamp.

Lazarsfeld, P. F. (1959). Psychology: A study of science. In S. Koch (Ed.), (Vol. 3). McGraw-Hill, New York.

Lehmann, E. L. (1997). *Testing statistical hypotheses.* New York: John Wiley & Sons, Inc.

Leichsenring, F. (1999a). Primitive defense mechanisms in schizophrenics and borderline patients. *Journal of Nervous and Mental Disorders*, *4*(187), 229–236.

Leichsenring, F. (1999b). Splitting: an empirical study. *Bulletin of the Menninger Clinic*, *63*(4), 520–537.

Link, J. (1999). *Versuch über den Normalismus. Wie Normalität produziert wird.* Westdeutscher Verlag: Opladen.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analyses with missing data.* New York: John Wiley & Sons, Inc.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories for mental test scores.* Addison-Wesley Publishing Company.

Maassen, G. H. (1998). The reliability weighted measure of reliable change as an indicator of reliable change. *Kwantitatieve Methoden*, *19*(58), 29–40.

Maassen, G. H. (2000a). Kelly's formula as a basis for the assessment of reliable change. *Psychometrika*, *65*(2), 187–197.

Maassen, G. H. (2000b). Principles of defining reliable change indices. *Journal of Clinical and Experimental Neuropsychology*, *5*, 622–632.

Malewski, P., & Dillmann, M. (1997). *Reliabilität und Standardmeßfehler des EDI (Eating Disorder Inventory).* (Tätigkeitbericht der Forschungsstelle für Psychotherapie)

Malewski, P., & Oehlschlägel, J. (2000). Gnu's S: Graphische und statistische Datenanalyse mit R. *Linux Magazin*, *3*(2), 124–127.

Meermann, R., & Vandereycken, W. (1987). *Therapie der Magersucht und Bulimia nervosa. Ein klinischer Leitfaden für den Praktiker.* Berlin: Walter de Gruyter.

Meyer, A., Richter, R., Graf, K. G. V., Schulenburg, J., & Schulte, B. (1991). *Forschungsgutachten zu Fragen eines Psychotherapeutengesetzes.* Bonn/Bad Godesberg: Gesundheitsministerium.

Mises, R. V. (1919). Wahrscheinlichkeitsrechung und Ihre Anwendung in der Statistik und Theoretischen Physik. *Zeitschrift für Mathematik*, *4*, 1–97.

Mossman, D., & Somoza, E. (1989). Maximazing diagnostic information from the dexamethasone suppression test. *Archive of General Psychiatry*, *46*, 653–667.

Nunnally, J. C., & Kotsch, W. E. (1983). Studies of individual subjects: logic and methods of analysis. *British Journal of Clinical Psychology*, *22*(83–93).

Oehlschlägel-Akiyoshi, J., Malewski, P., & Mahon, J. (1999). How to define anorectic weight? *European Eating Disorders Review*, *7*(5), 321–333.

Piaget, J., & Inhelder, B. (1986). *Die Psychologie des Kindes.* München: DTV.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed effect models in s and s-plus.* New York: Springer.

Reve, K. van het, & Busse, G. (1994). Reves Vermutung. In K. van het Reve (Ed.), *Dr. Freud and Sherlock Holmes* (pp. 140–151). Frankfurt am Main: Suhrkamp.

Rogosa, D. (1995). Myths and methods: ‚myths about longitudinal research' plus supplemental questions. In J. M. Gottman (Ed.), *The analysis of change* (pp. 3–66). New Jersey: Lawrence Erlbaum Associates.

Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement*, *4*, 335–343.

Rogosa, D. R., & Willett, J. B. (1985). Understanding correlates of change by modelling individual differences in growth. *Psychometrica*, *50*, 203–228.

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology*, *1*, 43-46.

Rudolf, G., Buchheim, P., Ehlers, W., Küchenhoff, A., J. Muhs, Pouget-Schors, D., Rüger, U., Seidler, G., & Schwarz, F. (1995). Struktur und strukturelle Störung. *Zeitschrift für psychosomatische Medizin*, *41*, 197–212.

Schmidt, S. J. (1987). Der Diskurs des radikalen Konstruktivismus. In S. J. Schmidt (Ed.), (pp. 11–88). Frankfurt am Main: Suhrkamp.

Schmitz, N. (1997). Klinische und Statistische Signifikanz - Diskutiert am Beispiel der Symptom Check Liste (SCL 90-R). *Diagnostica*, 80–96.

Speer, D. C. (1992). Clinically significant change: Jacobson & truax revisited. *Journal of Conculting and Clinical Psychology*, *60*, 402–408.

Spiegelhalter, D. J., Freedman, L. S., & Parmar, M. K. B. (1993). Applying bayesian ideas in drug development and clinical trials. *Statistics in Medicine*, *12*, 1501–1511.

Stegmüller, W. (1973). *Personelle und Statistische Wahrscheinlichkeit.* New York: Springer.

Steyer, R. (1984). Conditional expectations: An introduction to the concept and its application in empirical science. *Trierer Psychologische Berichte*, *11*(3), 1–24.

Steyer, R. (1992). *Theorie kausaler Regressionsmodelle.* Frankfurt am Main: Gustav Fischer.

Steyer, R., Ferring, D., & Schmitt. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment*, *8*, 70–98.

Steyer, R., & Schmitt, M. J. (1998). Basic concepts of latent state-trait theory. In *Consistency and specifity: Latent state-trait models in differencial psychology* (pp. 1–20). Rolf Steyer and H. Gräser and K. F. Widmann.

Sullivan, L. M., Dukes, K. A., & Losina, E. (1999). Tutorial in biostatistics an introduction to hierarchical linear modelling. *Statistics in Medicine*, *18*, 855-888.

Sutcliffe, J. P. (1965). A probability model for errors of classification general considerations. *Psychometrica*, *30*, 73–96.

Tack, W. H. (1986). Reliabilitäts- und Effektfunktionen - ein Ansatz zur Zuverlässigkeit von Messwertänderungen. *Diagnostica*, *32*, 48–63.

Thiel, A., Jacobi, C., Horstmann, S., Paul, T., Nutzinger, D. O., & Schüßler, G. (1997). Eine deutschsprachige Version des Eating Diasorder Inventory EDI-2. *Psychotherapie, Psychosomatik und Medizinische Psychologie*, *47*, 365–376.

Thiel, A., & Paul, T. (1988). Entwicklung einer deutschsprachigen Version des Eating-Disorder-Inventory (EDI). *Zeitschrift für Differenzielle und Diagnostische Psychologie*, *9*, 267–278.

Venables, W. N., & Ripley, B. D. (1999). *Modern applied statistics with s-plus.* New York: Springer.

von Wietersheim, J., Malewski, P., Jäger, B., Köpp, W., Gitzinger, I., Köhler, P., & Grabhorn, R. (2001). Der Einfluss von stationärer psychodynamischer Psychotherapie auf Persönlichkeitsmerkmale von Patienten mit Anorexia nervosa und Bulimia nervosa–Ergebnisse der multizentrischen Essstörungsstudie. *Zeitschrift für Psychosomatische Medizin und Psychotherapie*, *47*, 366–379.

Wellek, S. (1994). *Statistische Methoden zum Nachweis von Äquivalenz.* Gustav Fischer: Stuttgart.

Willett, J. B. (1988). Questions and answers in the measurement of change. In E. Rotzkopf (Ed.), *Review of research in education* (Vol. 15, pp. 345–422). Washington: American Educational Research Association.

Williams, R. H., & Zimmermann, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement*, *20*, 59–69.

Zegers, F. E., & Hafkenscheid, A. (1994). The ultimate reliable change index: An alternative to hageman & arrindell approach. *Heymans Bulletin*, *HB-94-1154-EX*.

Zimmermann, D. W., & Williams, R. H. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement*, *19*, 149–154.

Žižek, S. (2001). *Die Tücke des Subjekts.* Frankfurt am Main: Suhrkamp.

Žižek, S. (2003). *Die Puppe und der Zwerg. Das Christentum zwischen Perversion und Subversion.* Frankfurt am Main: Suhrkamp.

A Anhang

A.1 Anhang zur Stichprobenbeschreibung

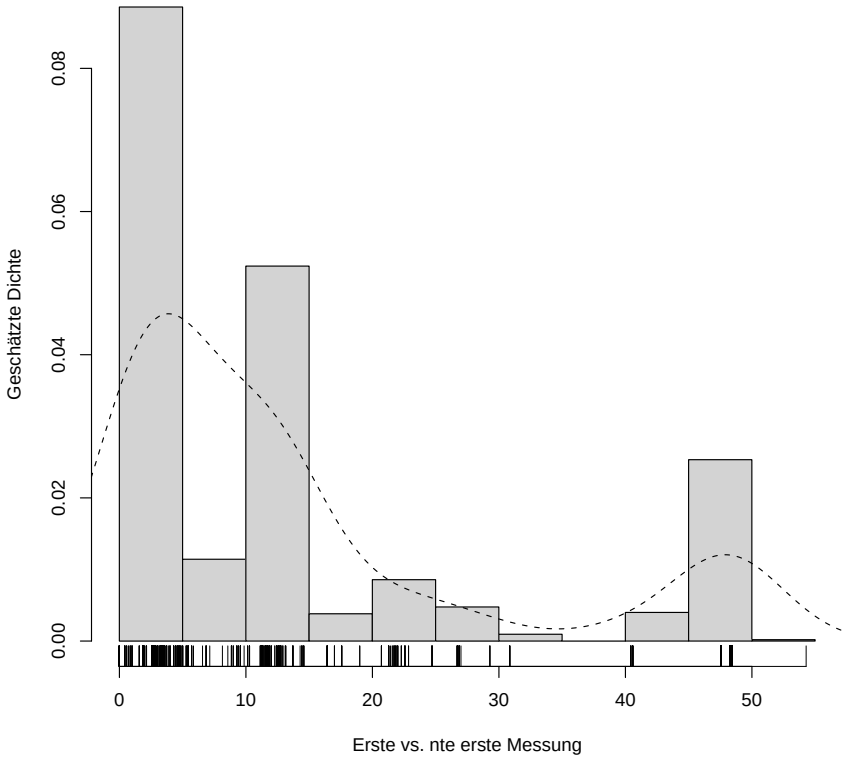

Abbildung A.1: Verteilung der eingegangenen Bögen relativ zum ersten Bogen.

A.2 Die Items des EDI-Münster

Die Items des EDI

Nr	Skala	Item
1	Drive for Thinness	Ich esse Süßigkeiten und Kohlenhydrate, ohne dabei nervös zu sein
2	Body Dissatisfaction	Ich empfinde meinen Bauch als zu dick
3	Maturity Fears	Ich wünschte, ich könnte zurückkehren in die Sicherheit meiner Kindheit
4	Bulimia	Ich esse, wenn ich mich durcheinander fühle
5	Bulimia	Ich stopfe mich mit Speisen voll
6	Maturity Fears	Ich wünschte, ich wäre jünger
7	Drive for Thinness	Ich denke über Diäten nach
8	Interoceptive Awareness	Ich bekomme Angst, wenn meine Gefühle zu stark werden
9	Body Dissatisfaction	Ich empfinde meine Oberschenkel als zu dick
10	Ineffectiveness	Ich fühle mich unfähig als Mensch
11	Drive for Thinness	Ich fühle mich schuldig, wenn ich mich überessen habe
12	Body Dissatisfaction	Ich glaube, dass mein Bauch gerade die richtige Größe hat
13	Perfectionism	In meiner Familie zählen nur hervorragende Leistungen
14	Maturity Fears	Die glücklichste Zeit im Leben ist die Kindheit
15	Interpersonal Distrust	Ich zeige offen meine Gefühle
16	Drive for Thinness	Ich habe Angst davor, zuzunehmen
17	Interpersonal Distrust	Ich vertraue anderen
18	Ineffectiveness	Ich fühle mich allein in der Welt
19	Body Dissatisfaction	Ich bin mit der Gestalt meines Körpers zufrieden
20	Ineffectiveness	Im allgemeinen habe ich das Gefühl, mein Leben unter Kontrolle zu haben
21	Interoceptive Awareness	Ich bin oft verwirrt über meine wahren Gefühle
22	Maturity Fears	Ich wäre lieber ein Erwachsener als ein Kind
23	Interpersonal Distrust	Es ist leicht für mich, mit anderen zu verkehren/reden
24	Ineffectiveness	Ich wünschte, ich wäre jemand anderer
25	Drive for Thinness	Ich übertreibe die Bedeutung des Körpergewichts

Liste der Items des EDI (Fortsetzung)

Nr	Skala	Item
26	Interoceptive Awareness	Ich kann meine Gefühle klar voneinander unterscheiden
27	Ineffectiveness	Ich fühle mich unzulänglich
28	Bulimia	Ich hatte schon Essanfälle, bei denen ich das Gefühl hatte, nicht mit dem Essen aufhören zu können
29	Perfectionism	Als Kind habe ich es immer angestrengt versucht zu vermeiden, meine Eltern und Lehrer zu enttäuschen
30	Interpersonal Distrust	Ich habe enge zwischenmenschliche Beziehungen
31	Body Dissatisfaction	Ich mag die Form meines Gesäßes
32	Drive for Thinness	Der Wunsch, dünner zu sein, nimmt mich geistig voll in Anspruch
33	Interoceptive Awareness	Ich weiß nicht, was in mir selbst vorgeht
34	Interpersonal Distrust	Ich habe Schwierigkeiten, anderen meine Gefühle zu zeigen
35	Maturity Fears	Die Anforderungen der Erwachsenenwelt sind zu hoch
36	Perfectionism	Ich hasse es, nicht der/die Beste zu sein
37	Ineffectiveness	Ich fühle mich in mir geborgen/bin mir meiner selbst bewußt
38	Bulimia	Ich beschäftige mich gedanklich mit Essanfällen
39	Maturity Fears	Ich bin froh, dass ich kein Kind mehr bin
40	Interoceptive Awareness	Ich weiß oft nicht, ob ich hungrig bin oder nicht
41	Ineffectiveness	Ich habe keine besonders gute Meinung von mir selbst
42	Ineffectiveness	Ich glaube, dass ich meine Ziele erreichen kann
43	Perfectionism	Meine Eltern haben hervorragende Leistungen von mir erwartet
44	Interoceptive Awareness	Ich habe Angst, dass meine Gefühle außer Kontrolle geraten
45	Body Dissatisfaction	Ich empfinde meine Hüften als zu breit
46	Bulimia	Vor anderen esse ich gemäßigt und stopfe mich erst dann voll, wenn ich wieder alleine bin
47	Interoceptive Awareness	Ich fühle mich schon nach einer kleinen Mahlzeit aufgequollen

Liste der Items des EDI (Fortsetzung)

Nr	Skala	Item
48	Maturity Fears	Ich glaube, dass Menschen am glücklichsten sind, wenn sie Kinder sind
49	Drive for Thinness	Wenn ich ein Pfund zunehme, habe ich Angst
50	Ineffectiveness	Ich glaube, dass ich ein wertvoller Mensch bin
51	Interoceptive Awareness	Wenn ich mich durcheinander fühle, weiß ich nicht, ob ich traurig, ängstlich oder wütend bin
52	Perfectionism	Ich habe das Gefühl, ich erledige Dinge entweder perfekt oder gar nicht
53	Bulimia	Ich denke daran zu erbrechen, um Gewicht zu verlieren
54	Interpersonal Distrust	Ich muss andere Menschen immer in einer gewissen Distanz halten/fühle mich unwohl, wenn jemand versucht, mir zu nahe zu kommen
55	Body Dissatisfaction	Ich glaube, dass meine Oberschenkel gerade die richtige Form haben
56	Ineffectiveness	Ich fühle mich innerlich leer
57	Interpersonal Distrust	Ich kann über persönliche Gedanken und Gefühle sprechen
58	Maturity Fears	Die besten Jahre im Leben sind die, wenn man erwachsen wird
59	Body Dissatisfaction	Ich empfinde mein Gesäß als zu breit
60	Interoceptive Awareness	Ich habe Gefühle, die ich nicht richtig einordnen kann
61	Bulimia	Ich esse oder trinke heimlich
62	Body Dissatisfaction	Ich bin zufrieden mit der Form meiner Hüften
63	Perfectionism	Ich habe sehr hohe Maßstäbe
64	Interoceptive Awareness	Wenn ich mich durcheinander fühle, habe ich Angst davor, dass ich anfangen könnte zu essen

Skala	Alpha	RetestKorrelation1.2	Intercept
Gesamt	0.942	0.918	0.937
Drive for Thinness	0.866	0.816	0.824
Bulimia	0.860	0.829	0.800
Body Dissatisfaction	0.945	0.933	0.935
Ineffectiveness	0.901	0.878	0.839
Perfectionism	0.781	0.814	0.834
Interpersonal Distrust	0.825	0.809	0.770
Interoceptive Awareness	0.858	0.833	0.829
Maturity Fears	0.753	0.616	0.764

Tabelle A.2: Tabelle mit den Realibilitätskoeffizienten des EDI (Test-Retest, Chronbachs Alpha, Steigung)

A.3 Reliabilitätskoeffizienten des EDI

A.4 RC-Änderungen: keine Änderung und Verschlechterungen

Keine Veränderung: Gewichtete Berechungsweise

Skala	Mit Zeit	Mit Intercept	mit Chron. Alpha	mit Retestr.
Gesamt	54.28	53.11	44.28	54.10
Drive for Thinness	51.02	51.02	50.04	50.67
Bulimia	49.42	46.76	41.26	46.76
Body Dissatisfaction	54.27	51.90	43.74	50.95
Ineffectiveness	65.04	65.04	74.00	65.04
Perfectionism	69.48	66.25	66.16	66.16
Interpersonal Distrust	60.31	51.13	51.13	51.13
Interoceptive Awareness	74.58	69.16	70.19	85.14
Maturity Fears	37.44	32.31	31.85	34.93

Verschlechterung: Klassische Berechungsweise

Skala	Mit Zeit	Mit Intercept	mit Chron. Alpha	mit Retestr.
Gesamt	1.26	1.44	2.43	1.44
Drive for Thinness	0.27	0.27	0.89	0.53
Bulimia	3.19	3.82	3.82	3.82
Body Dissatisfaction	0.85	1.04	1.80	1.80
Ineffectiveness	2.93	2.93	2.93	2.93
Perfectionism	1.53	1.53	2.87	2.87
Interpersonal Distrust	1.26	2.34	2.61	2.34
Interoceptive Awareness	1.03	1.03	1.03	0.65
Maturity Fears	1.14	2.17	2.28	1.48

Verschlechterung: Gewichtete Berechungsweise

Skala	Mit Zeit	Mit Intercept	mit Chron. Alpha	mit Retestr.
Gesamt	1.35	2.43	3.51	1.44
Drive for Thinness	0.53	0.53	1.51	0.89
Bulimia	5.77	8.25	8.34	8.25
Body Dissatisfaction	1.42	1.80	3.70	2.56
Ineffectiveness	6.57	6.57	6.57	6.57
Perfectionism	2.87	5.03	5.12	5.12
Interpersonal Distrust	1.26	2.61	2.61	2.61
Interoceptive Awareness	1.96	2.99	1.96	0.65
Maturity Fears	2.17	3.20	3.65	2.74

Verschlechterung: klassische Berechungsweise

Skala	Mit Zeit	Mit Intercept	mit Chron. Alpha	mit Retestr.
Gesamt	71.65	67.24	59.86	67.24
Drive for Thinness	73.87	73.87	62.13	68.09
Bulimia	68.68	63.62	63.62	63.62
Body Dissatisfaction	74.38	69.45	63.09	63.28
Ineffectiveness	88.73	88.73	88.82	88.73
Perfectionism	85.01	85.01	79.53	79.62
Interpersonal Distrust	75.97	68.14	64.09	68.14
Interoceptive Awareness	89.81	89.35	89.35	91.96
Maturity Fears	52.85	46.46	44.75	50.91

Überblick zu den unterstützenden Maßnahmen

Maßnahme	Häufigkeit
(Analytische) Psychotherapie, Einzel.	8624
Gestaltungstherapie, Gruppe	8397
Schwestern-/Pflegergespräche	7661
Wiegen	7488
Stationsvollversammlung	6195
(Analytische) Psychotherapie, Gruppe, ohne Co-TherapeutIn.	6105
(Medizinische) Visite	4342
(Analytische) Psychotherapie, Gruppe, mit Co-TherapeutIn.	3810
Medizinische Untersuchungen / Körperliche Untersuchungen	3484
Stationsgruppe	3168
Chefarztvisite	3153
Konzentrative Bewegungstherapie, Gruppe	2977
Autogenes Training	2964
Musiktherapie, Gruppe	2645
Körpertherapie (Gruppe)	2619
Oberarztvisite	2437
Stationsarztvisite	2344
Info-Gruppe: Ernährung (Grundumsatz etc.)	1992
Gespräche mit der Nachtbereitschaft	1784
Gymnastik	1672
Analytisch orientierte Einzelgespräche	1604
Maltherapie	1462
Medizinische Sprechstunde	1342
Entspannungstherapie (Gruppe)	1323
Behandlung durch Körperarzt	1287
Einzelgespräche mit TherapeutIn.(über 10 Min.)	1230
Essbegleitung für bulimische PatientInnen durch Schwester	1184
Frühgymnastik	1182
Körperwahrnehmungstherapie, Gruppe	1087
Morgendliches Treffen (Morgenrunde)	1082
Tonen / Töpfern	1049
Diätberatung	934
Massage/Bäder evt. mit Gymnastik	924
Kommunikative Bewegungstherapie	909
Beschäftigungstherapie	906
Ernährungsberatung	786
Info-Gruppe Psychologie (Grundbegriffe, Therapieformen)	782
Bewegungstherapie, Gruppe	696
Massage	657
Sportgruppe / Gymnastik	656

Überblick zu den unterstützenden Maßnahmen

Maßnahme	Häufigkeit
Entspannungstherapie nach Jacobsen(Gruppe)	649
Interaktionsübungen / Soziotherapie	647
Beratung durch SozialarbeiterIn (Gruppe)	639
Tanztherapie	614
Psychodrama, Gruppe	598
Partner- /FamiliengesprächeFamilien- /Ehetherapie	587
Familienskulptur	566
Gesprächsgruppe	512
Stationsbesprechung	494
Konzentrative Bewegungstherapie, Einzel.	490
Internistische Therapie	489
Atemtherapie (Gruppe)	474
Frühsport (Gruppe)	472
Beratung durch Sozialarbeiterin	449
Sporttherapie, Gruppe	447
Auf Sozial- und Berufsbereich themenzentrierte Gruppe	434
Märchentherapie	430
Übung sozialer Kompetenzen (Soziotherapie)	419
Psychologische Untersuchung (Standard-Interview)	410
Ergotherapie (handwerkliche Arbeitstherapie)	408
Fangopackungen	308
Essgruppe	308
Sozial(pädagogische) Beratung, Einzel.	293
Freies Gestalten	290
Psychotherapie-Visiten mit Behandlungsteam	287
Gemeinsame Mahlzeiten (mit Anwesenheit eines/einer PlegerIn)	286
Märchen-Malen	276
Schwunggymnastik	273
Themenzentrierte Gruppentherapie, mit Co-TherapeutIn	269
Gymnastik im Bewegungsbad	251
Gestaltungstherapie, Einzel.	244
Seidenmalen	241
Mal- und Gestaltungstherapie (Gruppe)	209
(Analytische)Psychotherapie,Gruppe geschlossen, ohne Co-Ther	203
Ernährungseinführung bei Aufnahme	202
Qi Gong Bewegungsübungen	201
Atem- und Leibtherapie (Einzel.)	195
Werkgruppe	183
Begrüßung zum Familienseminar	177
Abendgruppe	170
Therapeutische Großgruppe	168

Überblick zu den unterstützenden Maßnahmen

Maßnahme	Häufigkeit
Milieutherapie	167
Klinikvollversammlung (Gruppe)	166
Spezielle Gruppentherapie für Essgestörte	163
Wirbelsäulengymnastik (Gruppe)	156
Interaktionelle Gruppentherapie, ohne Co-TherapeutIn.	156
Beratung durch SozialarbeiterIn im Familienseminar in der Gr	156
Nachsorge-Einführung bei Aufnahme	156
Patientenvollversammlung (Komitee)(Gruppe)	154
Tanz und Bewegung (Gruppe)	154
Therapeutisches Malen	153
Einzelgespräch Schwester/Pfleger	153
Pädagogische Rollenspiele	152
Körperwahrnehmungstherapie, Einzel.	150
Medikamente	149
Schwimmen	147
Verhaltenstherapeutische Beratung	145
Katathymes Bilderleben	141
Krankengymnastik (Einzel.)	138
Bewegungsimprovisation (Gruppe)	137
Mal- und Spieltherapie	132
Vorstellungsgespräch beim Chefarzt	129
Zimmerrunde	126
Feldenkrais	124
Körperwahrnehmung / Eutonie (Gruppe)	122
Musikalische Selbstbesinnung und Körperwahrnehmungs-Gruppe	121
Literaturgruppe	120
Psychodynamische Therapie (Einzel.)	119
Rhythmik	118
Konzentrative Entspannung	118
Yoga, Autogenes Training	114
Stationsgruppe (gemeinschaftsorientiert)	113
Dienstvisite (täglich)	111
Krisentermine bei der/dem EinzeltherapeutIn	109
Analytische Suchtgruppe (Essstörungen)	104
Entspannungsgruppe	102
Körperwahrnehmungsgruppe	101
12-Schritte Gruppe	100
Internistische Diagnostik	98
Testpsychologische Untersuchung	96
Stationsgruppe (themenzentriert)	94
Leichte Spiele (Gruppe)	91

Überblick zu den unterstützenden Maßnahmen

Maßnahme	Häufigkeit
Musiktherapie, Einzel.	90
Gesprächsgruppe III (stützend)	84
Körpertherapie (Einzel.)	83
Teamvisite	83
Interaktionelle Gruppentherapie, mit Co-TherapeutIn.	80
Nottermine beim Pflegepersonal	79
Angsttraining	78
Tiefenpsychologische Anamnese	77
Ergometertraining	76
OA-Gespräche (Therapiebilanzgespräche)	74
Bewegungstherapie, Einzel.	73
Milieugruppe	72
Klausur (Einzelgespräche mit TherapeutIn)	68
Therapeutische Vertragsgruppe	67
Kalorienreiche Zusatznahrung	64
Info-Gruppe Nachsorge	63
Kunsttherapie, Einzel.	63
(Anal.) Psychotherapie, Gruppe, mit Co-Therapeut	56
Katathymes Bilderleben (Einzel.)	55
Pantomime	53
Esssucht Vertragsgruppe	53
Termine mit dem Oberarzt	52
Themenzentrierte Interaktion	51
Infogruppe	50
Feldenkrais orientierte Gruppe	46
Nachruhe nach dem Mittagessen	46
Termine mit Arzt vom Dienst	46
Familiengespräche	43
Bereitschaftsgruppe am Wochenende	43
Interaktionelle Gruppe	43
Analytische Maltherapie	42
Schwimmtherapie	41
Bogenschießen	40
Therapiegruppe für sexuell mißbrauchte Frauen	40
Yoga	39
Ehetherapie / Paartherapie	39
Themenzentrierte Suchtgruppe(Alkohol, Medikamente, Drogen)	39
Spezifische Betreuung (Zusatzernährung)	39
Sportgruppe/Körpersprache	36
Trampolingruppe	35
Katathymes Bilderleben (Gruppe mit Co-Leit.)	34

Überblick zu den unterstützenden Maßnahmen

Maßnahme	Häufigkeit
Zweitsicht und Wiedervorstellung	32
Entspannungstherapie (Einzel.)	31
Krankengymnastik (Gruppe)	31
Behandlung durch Konsiliararzt	29
Therapeutisches Gruppenmarathon	27
(Analytische) Psychotherapie, Gruppe, mit Co-TherapeutIn ('E	27
Sporttherapie, Einzel.	25
Reittherapie	25
Essmarathon (Gruppe)	25
(Analytische)Psychotherapie, Gruppe geschlossen, mit Co-Ther	24
Erwachsenenbildung (Gruppe)	24
Besprechen von Protokollen (z.B. Schmerztagebuch)	24
Atemtherapie (Einzel.)	23
Themenzentrierte Gruppentherapie, ohne Co-TherapeutIn	23
Psychodrama, Einzel.	22
Chefarztgespräch	22
Frauenspezifische Selbsterfahrungsgruppe	20
Inventurmarathon (Gruppe)	20
Bibliotherapie	20
Themenzentrierte Suchtgruppe(Sexsucht)	19
Sonstige medizin. Diagnostik (konsiliarärztliche Diagnostik)	19
Reflexzonenmassage	19
Labordiagnostik / apparative Diagnostik	19
Verhaltenstherapie (Einzel.)	18
Atem- und Leibtherapie (Gruppe)	17
Speckstein	16
Rezeptive Entspannungstherapie (Gruppe)	16
Kaufmännische Arbeitstherapie	15
Ausgang nur mit PflegerIn	14
Einführungsgruppe	14
Familientherapie	13
Psychologische Untersuchung	13
Theatergruppe	13
Wirbelsäulengymnastik (Einzel.)	12
Puppenspiel, Gruppe	12
Aufnahmegruppe	12
Gewichtsorientierte Ausgangsregelung	11
Physiotherapie, Einzel.	11
Chefarztverabschiedung	11
Schreibwerkstatt	11
Spezialbehandlung (z.B. physikalische Therapie)	11

Überblick zu den unterstützenden Maßnahmen

Maßnahme	Häufigkeit
Regulative Musiktherapie	10
Diät- und Ernährungsberatung	10
Arztsprechstunde	10
Photokurs	9
Verhaltenstherapie	9
Auseinandersetzungsübungen	9
Freies Gestalten (Einzel.)	7
Sprachtherapie	7
Männerspezifische Selbsterfahrungsgruppe	6
Beschäftigungstherapie (Einzel.)	6
Einzelgespräch bis 10 Min.	6
Puppen basteln	5
Familiengespräche mit CotherapeutIn	5
Übung sozialer Kompetenzen (Einzel.)	5
Therapeutisches Tonen	5
Gesundheitsdiskussion (Großgruppe)	5
Kunsttherapie	4
Kognitive Verhaltenstherapie (Einzel.)	4
Paargespräche ohne Co-Therap.	4
Angstkonfrontation	4
Skulpturgruppe	4
Stufenplan - Gewichtszunahme	4
Entspannungstherapie nach Jacobsen(Einzel.)	3
Familiengespräche ohne CotherapeutIn	3
Gemeinsame Mahlzeiten mit TherapeutIn	3
Paargespräche mit Co-Therap.	3
Partnergespräche	2
Therapie-Ausflug	2
Biofeedback	2
Angehörigengruppe	2
Eurythmie	1
Kognitive Verhaltenstherapie	1
Familienrekonstruktionsgruppe	1
Puppen basteln (Einzel.)	1
Gymnastik (Einzel.)	1
Vorgespräch (ambulant)	1
Gruppe für junge Erwachsene	1
Bulimiegruppe	1

Tabelle A.3: Anzahl der therapeutischen Maßnahmen.

Lebenslauf

seit 11.1998	Wissenschaftlicher Mitarbeiter an der Medizinischen Hochschule Hannover, Abteilung Psychosomatik und Psychotherapie (Leiter: Prof. F. Lamprecht).
02.1997–11.1998	Wissenschaftlicher Mitarbeiter an der Forschungsstelle für Psychotherapie Stuttgart (Leiter: Prof. H. Kächele). Mitarbeit im Projekt „Multizentische Studie zu Essstörungen “.
04.1998 – 10.1995	Studium der Psychologie (Diplom) an der TU-Braunschweig.
1984 – 1987	Fachgymnasium Wirtschaft in Osterode am Harz.
1980 – 1984	Realschule Röddenberg in Osterode am Harz.
1974 – 1978	Grundschule Dreilinden in Osterode am Harz.
11.06.1967	Geboren in Osterode am Harz. Eltern: Rudolf Malewski & Christel Malewski geb. Kikillus.